ote technical note techn

Human Factors Evaluation of Vocoders for Air Traffic Control Environments Phase I: Field Evaluation

James LaDue, Ph.D., SRC
Randy L. Sollenberger, Ph.D., ACT-530
Bill Belanger, PE, EPA
Annemarie Heinze, SRC

September 1997

DOT/FAA/CT-TN97/11

U.S. Department of Transportation
Federal Aviation Administration

William J. Hughes Technical Center
Atlantic City International Airport, NJ 08405

NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents or use thereof.

The United States Government does not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the objective of this report.

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/CT-TN97/11		

4. Title and Subtitle	5. Report Date
Human Factors Evaluation of Vocoders for Air Traffic Control Environments Phase I: Field Evaluation	September 1997
	6. Performing Organization Code ACT-530

7. Author(s)	8. Performing Organization Report No.
James La Due, Ph.D., SRC, Randy Sollenberger, Ph.D., ACT-530, Bill Belanger, PE, EPA, and Annemarie Heinze, SRC	DOT/FAA/CT-TN97/11

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Federal Aviation Administration William J. Hughes Technical Center Atlantic City International Airport, NJ 08405	11. Contract or Grant No. 94C-00042

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Federal Aviation Administration Communications & Infrastructure Branch William J. Hughes Technical Center Atlantic City International Airport, NJ 08405	Technical Note
	14. Sponsoring Agency Code ACT-330

15. Supplementary Notes

16. Abstract

Communication congestion is a major problem facing the air traffic control system. Vocoders offer a potential solution to this problem by compressing a digitized human speech signal to achieve low bandwidth voice transmissions. Air traffic controllers and pilots must find new systems usable and acceptable before the FAA authorizes implementation. This study compared the performance of two 4.8 kbps vocoders (designated as A and B) with the current analog radio system. Two hundred and seven current air traffic controllers participated in the study. Participants listened to recorded audio messages and provided written responses. The dependent measures included both subjective ratings and objective measures of intelligibility and acceptability. The research design controlled the independent measures of sex of speaker, background noise, and communication equipment. The results indicated that analog radio and vocoder B communications scored subjectively similar. Participants rated radio higher than vocoder B in intelligibility and vocoder B higher than radio in acceptability. They gave Vocoder A the lowest ratings using the subjective scales. An objective message completion test revealed that vocoder B was more intelligible than vocoder A. The results found no generally preferred sex of speaker for vocoder transmissions. There were no major effects of cockpit background noise on the communications.

17. Key Words	18. Distribution Statement
Human Factors Communications Air Traffic Control Vocoders Digitized speech Aeronautical communications	This document is available to the public through the National Technical Information Service, Springfield, Virginia, 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	60	

Form DOT F 1700.7 (8-72) **Reproduction of completed page authorized**

Acknowledgement

The authors would like to gratefully acknowledge the subject matter expertise provided by Randall Phillips of Cleveland Center and Patricia Horan of ACT-212. In addition, the authors would like to recognize the engineering support provided by ACT-330, especially James Eck, Edward Colemen, and John Petro.

Table of Contents

Table of Contents (cont.)

List of Illustrations

List of Illustrations (Cont.)

Figures Page

Tables Page

List of Illustrations (Cont.)

Tables Page

Executive Summary

This study evaluated the effectiveness of 4.8 kbps voice coders (vocoders) to determine the feasibility of future implementation into the Air Traffic Control (ATC) system. The implementation of vocoders into the ATC system would offer many advantages including an increase in the available number of ATC communication channels.

This study examined two vocoder models. The tests designed to evaluate the vocoders used subjective ratings and objective measures. Current controllers from Air Route Traffic Control Center, Air Traffic Control Tower, and Terminal Radar Control environments participated in this study. The general test format consisted of prerecorded audio messages with written responses by the participants. In the Subjective Ratings Test, the controllers rated both intelligibility and acceptability of the voice messages using 8-point scales. The test conditions included sex of speaker, background noise, and communication equipment. The Message Completion Test served as an objective measure of vocoder performance. In this test, the controllers wrote responses in three blanks for each presented message. Finally, an Audio Preference Test measured which vocoder was preferable. In this test, the participants selected the vocoder they would prefer to use in the field. All tests were double blind in that both participants and field researchers were not informed which vocoder was which and what tape sequences represented which conditions.

The results indicated that analog radio and vocoder B communications scored similarly, with radio higher than vocoder B in intelligibility and vocoder B higher than the radio in acceptability. Vocoder A rated as the least intelligible and least acceptable communications equipment. The Message Completion Test revealed that with the male speaker, vocoder B was more intelligible than vocoder A. However, controller accuracy on this test was near perfect and no meaningful comparisons arose between the vocoders and analog radio. The Audio Preference Test indicated that a clear preference existed for vocoder B over vocoder A. The results indicated no generally preferred sex of speaker for vocoder transmissions as well as the effects of jet, propeller, and helicopter aircraft background noises on the communications.

1. Introduction

1.1 Background

With the growing complexity of the current Air Traffic Control (ATC) system, the search to increase communications systems capacity is critical. Expanding the number of available channels in the ATC system from the current 25 kHz bandwidth spacing will help. Current voice coder (vocoder) digital technology offers a potential solution to this problem and several others in the VHF analog system.

Waveform coders can transfer a digital signal over a communication line and then reassemble the signal at the receiving end. These coders make no assumption about the coded signals and, as such, are subject to high bit rates (64 kbps) in the transfer of the information. Vocoders assume an explicit model of speech production in an attempt to produce a signal that sounds like original speech, whether or not the time waveform resembles the original. The result is the production of intelligible speech at very low bit rates. However, the speech produced can sound synthetic and can be subject to intelligibility problems based on the assumptions made in the model used to characterize the speech.

The use of digital communication offers several advantages over current analog communication systems. The vocoder can achieve these advantages in a far more bandwidth-efficient manner. By using vocoder technology, the voice signaling can reduce to 4.8 kbps, allowing four channels to be assigned within the 25 kHz bandwidth currently used by a single analog channel. These additional channels can be used for voice or data communications. The use of digital technology also provides a means of secure communications and offers potential solutions to the problems of stuck microphones and "stepped on" communications.

1.2 Problem Statement

The goal of this study was to evaluate the effectiveness of 4.8 kbps vocoders to deliver en route and terminal voice messages. There are many factors that affect the quality of vocoder transmissions. This study addressed the multiple effects of sex of speaker, background noise, and specific vocoder equipment. This study evaluates two vocoders, which differ in hardware and the speech coding algorithm applied.

1.3 Literature Review

Fike and Friend (1983) provide useful background into digital voice transmission techniques. One way to reduce the digital transmission rate with communications involving human speech is to use an algorithm to capture the speech in a narrow bandwidth. This is termed speech coding. Tremain and Collura (1988) described the coding techniques used in several different proprietary vocoder models. Because there is no universally accepted coding scheme, vocoder behavior varies with manufacturer and application requiring a thorough evaluation prior to successful implementation.

Pickens (1996) reviewed the four fundamental dimensions of vocoder evaluation: efficiency, delay, complexity, and performance. The first three of these may be expressed in objective terms, but vocoder performance is often a function of the opinions of the users. As such, subjective measures are often used to measure the speech quality of vocoder communications.

Panzer, Sharpley, and Voiers (1993) reviewed some subjective methods used for evaluating speech quality. The most difficult aspects of such studies lie in the quantification of properties of voice transmissions in which users would find significant subjective bias. Subjective ratings can be user specific and can often vary by user from test to test, a point emphasized by Voiers (1983). In addition, many of the perceived qualities of voice transmissions are heuristic and difficult to define requiring the use of trained listeners for evaluation. One such method is the Diagnostic Acceptability Measure (DAM) (Dynastat, 1995). This method combines a direct (isometric) and indirect (parametric) approach to acceptability evaluation. In addition to rating acceptability of a speech sample directly, listeners can indicate, independently, the extent to which various perceived qualities are present in the sample without regard to how they may affect acceptability.

Even with these significant difficulties involved with subjective evaluations, particularly with untrained listeners, there are some subjective methods in the literature for vocoder testing. Although the results of previous vocoder testing do not directly apply to present vocoders due to advances in the technology, the measurement methods used provide insight into techniques for subjective voice quality measurement.

Crowe (1988) used a subjective evaluation technique with untrained listeners based on a 5-point scale for rating the listening quality of speech communications for vocoder usage with the Aeronautical Satellite Service. Troll (1989) used a 5-point scale to rate intelligibility of vocoder communications. He also established a relative comparison between vocoder models with application to vocoder usage for ATC/pilot communications by satellite links. The test subjects for this evaluation were Air Traffic controllers. Child, Cleve, and Grable (1989) used air traffic controllers as participants and a 5-point subjective scale to rate both intelligibility and acceptability of vocoder equipment for ATC communications. For most vocoders tested, the intelligibility rated approximately 5% higher than the acceptability. This indicated that there were distinctions between the factors that influence vocoder intelligibility and those related to acceptability. Kemp, Sueda, and Tremain (1989) and Tremain, Kemp, Collura, and Kohler (1993) used the DAM previously mentioned with untrained listeners as well with success.

Although objective measures have not been widely used in the evaluation of the quality of voice communications, these methods have yielded significant information on message intelligibility. Sanders and McCormick (1987) reviewed several aspects of a speech transmission system that can affect intelligibility. These include frequency distortion, filtering, amplitude distortion, and modification of the time scale. Dynastat (1995) created an objective measure referred to as the Diagnostic Rhyme Test (DRT) to measure the effects of these forms of transmission distortion on intelligibility. This test consists of word pairs separated by a single consonant sound. Listeners select one out of two words heard. More study-specific methods such as work done by Dehel, Grable, and Child (1989) used an objective evaluation for vocoder assessment including

an open word test, the DRT, and a number recognition test. Troll (1989) evaluated vocoder communications for satellite digital communication links using readback tests for whole and split sentences. The study found that the memory of the controller test subjects became more important as the length of the clearance increased. However, the method worked very well for split tests in which the test participants read back only key parts of a clearance, such as a radio frequency or call sign.

The effects of context are another primary issue in measuring the intelligibility of speech communication systems. The systems seldom will operate in an informational void. Sanders and McCormick (1987) note that intelligibility is higher for sentences than for isolated words because the context supplies information. Intelligibility varies with the size of the vocabulary used. Miller, Heise, and Lichten (1951) studied the intelligibility of words from vocabularies of different sizes under varying signal-to-noise ratios. The percentage of words correctly recognized correlated strongly with the size of the vocabulary used and the signal-to-noise ratio. Voiers (1983) studied the relevance of context and concluded that the level of contextual support is not measurable. This led to measures of speech performance of a system that are totally free of context such as the DAM mentioned previously. However, a context-free environment is not universal especially in ATC communication evaluations (e.g., Dehel et al., 1989). Vocoder studies in which context yields support include Crowe (1988), Troll (1989), Child et al. (1989) and Grable (1990).

One important test condition for vocoder studies in an aviation environment involves studies of background noise and its effects on vocoder performance. Sanders and McCormick (1987) state that noise is the bane of speech intelligibility. A study done by Warren (1996) indicated that high noise levels are quite common in aviation environments, but noise canceling equipment is quite useful in alleviating the problem. However, Warren noted that any vocoder evaluation should include the harsh effects of background noise. For consideration of background noise level, Hart (1988) reviews several types of cockpit noise and provides usage levels as a benchmark for experimental comparison. The Federal Aviation Administration (1989) provides information concerning the relationship between flight crew cockpit voice communication and cockpit noise levels. The FAA recommends measuring cockpit noise using the "A" scale on a sound level meter. This choice of scale closely approximates the range of frequencies most prone to interfere with human speech.

There has been a great deal of research accomplished in audio communication within many contexts. The research and methods used provide support for the work reported here. However, the human factors research team had to develop new techniques to evaluate vocoders because the results required credibility with air traffic controllers.

2. Method

2.1 Experimental Design

2.1.1 Independent Variables

Three independent variables were manipulated in this study: type of voice communication equipment, sex of speaker, and background noise. The following paragraphs describe these variables.

Central to this study was the comparison of two 4.8 kbps vocoders from different manufacturers. The researchers referred to these vocoders simply as vocoder A and vocoder B and were not informed of the manufacturer's identity or any proprietary information.

While it seemed reasonable that the control would consist simply of non-vocoded speech, there were some considerations that led to another choice. In previous vocoder testing with ATC personnel, Dehel et al. (1989) presented the experimental participants with six vocoder models plus a clear signal as control. Approximately 90 % of the subjects rated the clear signal as "much better than current communications," which indicated that the control was not representative of typical analog ATC radio communications. The validity of the previously used control came into question. The use of such a superior audio presentation would bias the participants against the digitized vocoded speech because it provides an unrealistic benchmark for comparison. For this test, the control was an analog signal that sounded similar to typical ATC communications using aircraft radios. Attenuation of the signal produced a signal-to-noise ratio typical of ATC environments. The signal-to-noise ratio was set by adjusting the signal strength to a level corresponding to 50% of the effective range guaranteed within the service volume of ATC antennas. This control produced a realistic simulation of actual analog ATC communications. The researchers did not pass the vocoder speech through the analog radio path. This was because the digitized speech does not degrade in the same way as radio signal-to-noise ratio. In a digital transmission, radio "static" is not audible to the listener's ear. Instead, it manifests itself as errors in the digital bits decoded by the receiving equipment. A fixed Bit Error Rate (BER) integrated these "bit errors" in this study.

The sex of the speaker was another independent variable included in the study. The significance of this variable was primarily in the reduction of bandwidth that accompanies vocoder operation. The compression algorithms of the vocoders may result in unequal communication quality when considering the different pitch and tone in which male and female voices operate. It was an aim of this study to ascertain if these differences created any degradation in communication intelligibility or quality.

Dehel et al. (1989) revealed a sensitivity of earlier vocoder designs to background noise in the ATC application environment. Although the present vocoders are later realizations of vocoder technology, the noisy application environment demanded assessing these effects once again.

For this study, there were three general categories of background noise. The first was a propeller aircraft in cruise flight. This noise has application to communications originating from general

aviation aircraft and small commercial aircraft. The second was that of a jet. This noise condition referred to background sound heard from the pilot's perspective of a passenger turbojet aircraft in cruise flight. The third was that of a turbine helicopter in cruise flight. The standard for comparison was a quiet environment with no background noise.

The selected background noise levels were similar to the average noise levels of aircraft in each of the classes examined. For propeller aircraft, the literature revealed measurements of 15 light single-engine airplanes (Tobias, 1968a), 11 light twin-engine airplanes (Tobias, 1968b), and 3 FAA-operated propeller aircraft (Rodgers, 1995). Noise levels ranged from 90-92 dB[A] for the singles, 88 dB[A] for the light twins, and 85 dB[A], 87 dB[A], and 91 dB[A] in the FAA-operated aircraft. Based on this information, the propeller background noise was set at 90 dB[A], which is representative of most propeller aircraft.

For jet aircraft, there appeared to be a significant spread in the noise level in the cockpit. Three FAA-operated aircraft gave noise levels of 78 dB[A], 92 dB[A], and 93 dB[A] (Rodgers, 1995). Cockpit noise measured for this study in a commercial DC-9 yielded an intensity level of 79 dB[A]. The slightly greater than 10 dB spread represents a factor of 10 in the acoustic power level in the cockpit. To include the jets with higher cockpit noise, the jet cockpit noise background was recorded at 90 dB[A] with the understanding that it represents a near-critical level in civilian aircraft.

Turbine helicopters fell into two broad categories, civilian and military. The military types appear to have a significantly higher cockpit noise level than civilian helicopters. Background noise in military helicopters ranges from 90 dB[A] in the AH-1 and OH58D to 115 dB[A] in the CH47C (Hart, 1988). For this work, cockpit noise was measured in a Sikorsky S-76 helicopter and a Bell Long Ranger. The average sound level in normal cruise flight was 90 dB[A] in the Sikorsky and 90 dB[A] in the Long Ranger. Based on this information and the literature, the researchers chose a turbine helicopter background noise level of 90 dB[A]. The exclusion of the higher noise levels of military helicopters stemmed from the ability of their pilots to use microphones in face masks and other means to suppress cockpit noise. Civilian pilots do not use these methods.

The intensity levels were therefore 90 dB[A] across all categories. This constant level simplified the experimental conditions in eliminating the comparative effect of intensity level yielding a more precise analysis of the results. Table 1 presents a summary of the sources of the cockpit background noises.

2.1.2 Summary of Independent Variables

Table 2 indicates a summary of the relevant independent variables for this study. The independent variables and their associated levels yielded 24 test conditions for the evaluation. Several additional variables were considered for this study but were not included as explained in the following paragraphs.

Table 1. Description of Aircraft Background Noises Used in Audio Recordings

Background Noise	*Category*	*Specific Source*	*Intensity Level*
Propeller aircraft	Singles, light twins, turboprops	Rockwell Twin Aero Commander	90 dB[A]
Jet aircraft	Large commercial jets, small corporate jets	Boeing 727	90 dB[A]
Helicopters	Civilian helicopters	Sikorsky SK-76	90 dB[A]

Table 2. Independent Variables

Independent Variable	*Number*	*Levels*
Equipment	3	Vocoder A, Vocoder B, Analog Radio
Aircraft Background Noise	4	Jet, Propeller, Helicopter, None
Sex of Speaker	2	Male, Female
Total Test Conditions	24	

Bit Error Rate (BER) is the fraction of the transmitted bits in a digital system that the receiving radio does not correctly interpret. The standard for the minimum transmission BER to be used with vocoders is currently undefined. In this study, the BER was a fixed value of 10^{-3}. This value was chosen because it has been the standard for comparison in previous vocoder testing. The inclusion of varying BERs would cause a significant increase in the number of test conditions and would unduly increase the complexity of the study.

Another variable that could potentially impact vocoder performance is the accent of the speakers. An adequate investigation of speaker accents would require an enormous research effort and is beyond the scope of this study. One male and one female controller without strong regional accents were selected as the voices on the recorded messages.

Distinctions in speech rate between different pilots and controllers are quite common. However, due to the large amount of test conditions deemed essential to include as variables in this study, including speech rate did not conform to constraints arising from practical considerations. The speech rate used by both the male and female controllers was a function of their experience controlling traffic. The controllers attempted to speak at a consistent rate so there were little or no differences in speech rate during the recording of the stimulus phrases.

2.2 Test Format

The test format consisted of the following three parts:

a. Subjective Ratings Test

b. Message Completion Test

c. Audio Preference Test

The participants recorded each measure by written responses. These test measures are described in the following sections.

2.2.1 Subjective Ratings Test

This test consisted of the participants rating the intelligibility and acceptability of the audio messages on an 8-point scale. The messages consisted of clearances, pilot readbacks, or pilot requests; none exceeded 30 syllables.

In the Subjective Ratings Test, the controllers evaluated three measures of intelligibility and acceptability for each test condition. The average of these three ratings became the overall score for that test condition. Three replications gave some variability between the messages presented for each condition in addition to adding stability to the individuals' scores.

The intelligibility rating was defined as the ability to understand the spoken message. The poles of the 8-point scale contained the following anchors:

1	*poor*	Could not understand anything that was said during the transmission.
8	*excellent*	Understood everything that was relayed during the transmission precisely.

The definition of acceptability was gauged by the

a. quality of the message (i.e., annoying, pleasant),

b. effort required to understand the message (i.e. easy, burdensome), and

c. potential influence of the background noise (i.e., buzzing, hissing).

The test participants rated acceptability on an 8-point scale with the following anchors:

1	*poor*	Would be terribly annoying, frustrating, or unpleasant to hear.
8	*excellent*	Excellent signal quality, a clear signal that would be pleasant to hear.

2.2.2 Message Completion Test

This section of the evaluation is composed of the objective assessment of the vocoders. This test consisted of a series of messages delivered to the test participants by audio tape. Each message corresponded to a written phrase in the evaluation booklet by number. Each written message in the booklet contained three omissions. The test participants completed the blanks by writing the missing portions of the message. The blanks omitted contained information that the participants could not determine from the surrounding context of the message such as an aircraft call sign, frequency, or location.

The participants completed three messages per test condition, each message containing three blanks. With three blanks completed per message, nine objective evaluations of intelligibility were collected for each test condition presented to participants. An overall score from zero to nine was applied to each test condition based on the number of responses correctly filled.

To prevent the participants from becoming more adept at the test by reading ahead, the messages were not presented in the same order as they appeared on paper. Rather, on each page, the recording queued a message number followed by the message in a random order. When all applicable messages were completed on a page, the recording narrator asked the participants to turn the page.

2.2.3 Audio Preference Test

This test consisted of each test participant hearing the same message under the same test conditions from vocoder A and vocoder B. The test participants assessed which "audio format" they would prefer to use in actual field conditions. This test was to serve as an absolute measure to eliminate one of the vocoders from future phases of the project. There was a total of eight conditions for this test using the independent variables listed in Table 3. To present a fair comparison, each message was greater than 40 syllables and designed to include a large variety of voice sounds.

Table 3. Independent Variables for the Audio Preference Test

Independent Variable	*Number*	*Levels*
Aircraft Background Noise	4	Jet, Propeller, Helicopter, None
Sex of Speaker	2	Male, Female
Total Test Conditions	8	

The participants answered questions about which vocoder they would prefer and were asked to communicate the rationale for their choice under each test condition. The participants wrote a brief description on what aspects of what they heard led to their preference.

2.2.4 Double-Blind Control

All evaluations were conducted double blind. Neither the participants nor field researchers knew which vocoder was which and what tape segments represented what conditions. The primary reason for this double-blind aspect was to increase the validity of the results.

2.2.5 Message Context

Pilot-controller communications served as the context for all messages in all tests. The researchers insured there were no significant departures from ATC phraseology in the recorded messages. Controllers and pilots commonly make departures from this glossary in the field, which could increase vocabulary and make intelligibility more difficult. However, most departures are limited in scope to a particular situation or region. To insure a conservative approach, the message completion portion of the test provided an objective measure of intelligibility lacking contextual support. The blanks completed were often aircraft call signs, frequencies, or headings that the participants could not derive from the surrounding text.

A special consideration insured that the background noise corresponded to the message context. The researchers wanted to prevent a situation such as a heavy jet clearance being presented with

helicopter background noise, which could confuse the participants and create bias. All messages were, therefore, realistically comparable with those used in the field.

2.2.6 Test Length and Distribution

The research team decided that 90 minutes was a maximum time that could be asked of volunteer participants. Preliminary investigations indicated that each participant could not complete all test conditions with replications and remain under the time limit. To avoid this difficulty, the researchers split the Subjective Ratings Test and the Message Completion Test at the sex of speaker condition whereby each participant evaluated voice messages from a single sex of speaker for both tests. Completed tests from two participants evaluated all test conditions. Table 4 presents a summary of the test distribution.

The first two messages of the Subjective Ratings Test and Message Completion Test were practice for the participants. The average total time remained below the target limit including a 10-minute break and planned instructions, and rarely exceeded 80 minutes.

Table 4. Test Distribution

Test	*Total Test Conditions*	*Conditions Evaluated per Test Participant*	*Number of Messages*	*Test Time per Message*	*Approximate Test Time*
Subjective Ratings	24	12	3 per test condition + 2 Practice	60 seconds	38 minutes
Message Completion	24	12	3 per test condition + 2 practice	30 seconds	19 minutes
Vocoder Preference	8	8	1	60 seconds	8 minutes
				Total	1 hour, 5 minutes

2.3 Data Acquisition

2.3.1 Participants

The test participants consisted of 207 current air traffic controllers. The test participants were from Air Route Traffic Control Center (ARTCC), Air Traffic Control Tower (ATCT), and Terminal Radar Control (TRACON) environments. Strict adherence to all federal, union, and ethical guidelines on the use of human participants was upheld.

2.3.2 Experiment Staff

The personnel involved in the design and execution of this study included the Project Manager, who is a Human Factors Professional, two Human Factors Specialists, one Human Factors Engineer, a Supervisory Air Traffic Control Specialist (SATCS), and an Air Traffic Control Specialist (ATCS). The SATCS served as the Subject Matter Expert (SME) and created all

stimulus materials aided by the ATCS who was the female speaker on the test tapes. The two Human Factors Specialists were responsible for the test design, data analysis, and final report. The Human Factors Engineer was responsible for the hardware design, quality assurance of physical measurements, and troubleshooting. Several support engineers were also available for consultation. A Human Factors Specialist and the SME collected data in the field facilities.

2.3.3 Equipment and Procedures

The ATC test messages were recorded using the equipment described in Figure 1. The researchers selected the equipment and methods to realistically test the vocoders. The equipment and procedures below describe the addition of this background noise to the voice messages for the vocoder evaluation.

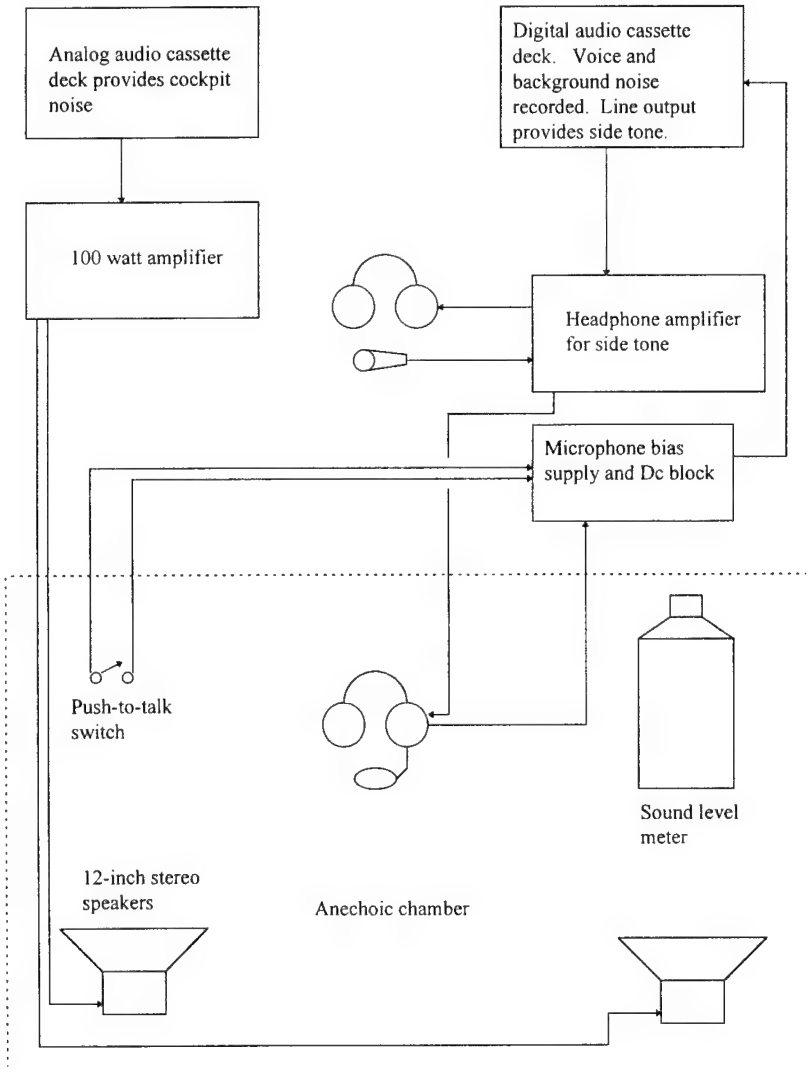

Figure 1. Source tape apparatus.

The simulation for constructing the source tapes used an anechoic chamber at the William J. Hughes Technical Center Research Development and Human Factors Laboratory. The size of the chamber is approximately 2 m x 1.3 m x 2 m high (6 ft x 4 ft x 6 ft). A Crown Macro-Tech 600 100-watt amplifier drove two Community 12-inch stereo speakers to provide the background noise. A Yamaha C300 professional stereo tape deck played the background noise tapes recorded in the cockpits into the amplifier. The playback level of the background noises closely matched the levels heard in the cockpits of the aircraft categories studied. Playback sound levels were monitored using a Radio Shack model 33-2055 sound level meter that was the same meter used to determine the cockpit noise levels. This meter was calibrated using an Occupational Safety and Health Administration (OSHA) standard noise source immediately before making the recordings.

To present a realistic test of the vocoder under the high-background noise volume levels, the researchers used a high-quality aviation headset with noise-canceling microphone to create the source tapes. This headset, representative of aviation headsets commonly in use, was a David Clark model H3330. In a test run, this equipment produced an audible voice in the presence of the highest background noise levels.

To use the aviation headset outside the cockpit environment, the researchers simulated the electronics of an aviation radio including the construction of a bias voltage supply to the microphone, a DC blocking capacitor, and a push-to-talk switch. Bias voltage was selected at 6 volts with a load resistance of 1.5 kΩ. A 22 μF DC blocking capacitor was used. These values produced the clearest possible speech through the microphone. The push-to-talk switch was a modified foot switch intended for control of electric tools. The switch was rewired as normally closed when the circuit broke at the push of the pedal. This switch connected in parallel with the microphone output after the blocking capacitor, which allowed the speakers to switch to on and off positions without any "clicks" or "pops" being transmitted to the recording device. Overall, this setup produced a realistic simulation of proposed vocoder operations. In addition, the simulation also realistically modeled the sounds normally heard by air traffic controllers and pilots as the background noise was present only when keying the microphone.

The side tone in the headset used in the sound chamber provided an additional element of realism. The output jack of the Sony TCD-D3 digital audio tape used to record the source tape provided this side-tone. The side tone is almost essential when using an aviation headset in a high-noise environment because it gives the person wearing the headset some feedback on what is transmitting through the microphone. Fike and Friend (1983) give a more complete examination of side tone effects.

Routing of the side tone was through a Tascam MH40 headphone amplifier to the aviation headset worn by the air traffic controllers who spoke the stimulus phrases. This amplifier also provided a second audio output for use by the recording system operator stationed outside the sound room that allowed monitoring of the audio quality during the creation of the tapes. A separate microphone was also attached to the headphone amplifier that allowed for headset switching between the recorder output and the operator's microphone. This switching allowed the operator to converse with the person in the sound room in the presence of the background

noise. This direct communication line allowed for quality assurance of the source material while recording the source tapes. This conversation was introduced at a point downstream of the recording device and was not recorded on the source tape.

Background noises were transferred from their original digital format to an analog audio tape. The analog tape was adequate for the playback of background noise because a large dynamic range was unnecessary. The 90-minute cassette contained 20 minutes each of jet, propeller, helicopter, and ATC noise. The sound level meter was placed in the sound booth in a position where it could be viewed through the booth window by the operator. In a test run, the researchers found that the sound level varied by only a few decibels if the meter was relocated throughout the booth. The results allowed placement of the sound level meter for operator convenience. With the sound level meter set for the appropriate scale, the background recording began. Adjustments of the background noise volume occurred until the sound level meter registered the correct noise intensity.

As the background noise played, the test speaker read from a script to produce the source tape. An approximate 5-second delay was allowed between the phrases. The digital audio tape recorder had an automatic feature that placed a marker on the tape if there was no sound input for 3 seconds. The automatic marking of the audio phrases allowed for quick referencing of particular phrases on playback. At this point, the source test material was pure in that there was no prefiltering of the audio spectrum except that imposed by the microphone itself and no degradation due to radio static. Eight tape segments were constructed using male and female speakers and three background noises plus a control with no background noise.

Eight individual source tape segments were initially produced for the Subjective Ratings Test and Message Completion Test. These tapes were processed through the communication equipment being tested and, subsequently, rerecorded using a Sony TCD digital audio tape recorder. The result was 24 source tape segments (two vocoders and the analog radio control). Each tape segment contained 20 voice phrases for the Subjective Ratings Test and 20 voice phrases for the Message Completion Test. Of these, 10 phrases were read by the male speaker and 10 were read by the female speaker. These tape segments were the sources for the development of the six master digital audio tapes.

Figure 2 illustrates the format for constructing the test tapes for the Subjective Ratings Test and the Message Completion Test. Researchers randomly assigned the source tape segments into three groups. The message groups were split by sex of speaker combinations in which one test version had a male speaker for the Subjective Ratings Test and a female speaker for the Message Completion Test. The other version had a female speaker for the Subjective Ratings Test and a male speaker for the Message Completion Test. In this way, each participant received one half of the total number of test conditions and had both male and female speakers included. The result was six test versions from the 24 source tape segments.

The test tape construction for the Audio Preference Test was straightforward. After constructing the master tape, which consisted of one message for each of the eight test conditions, the master

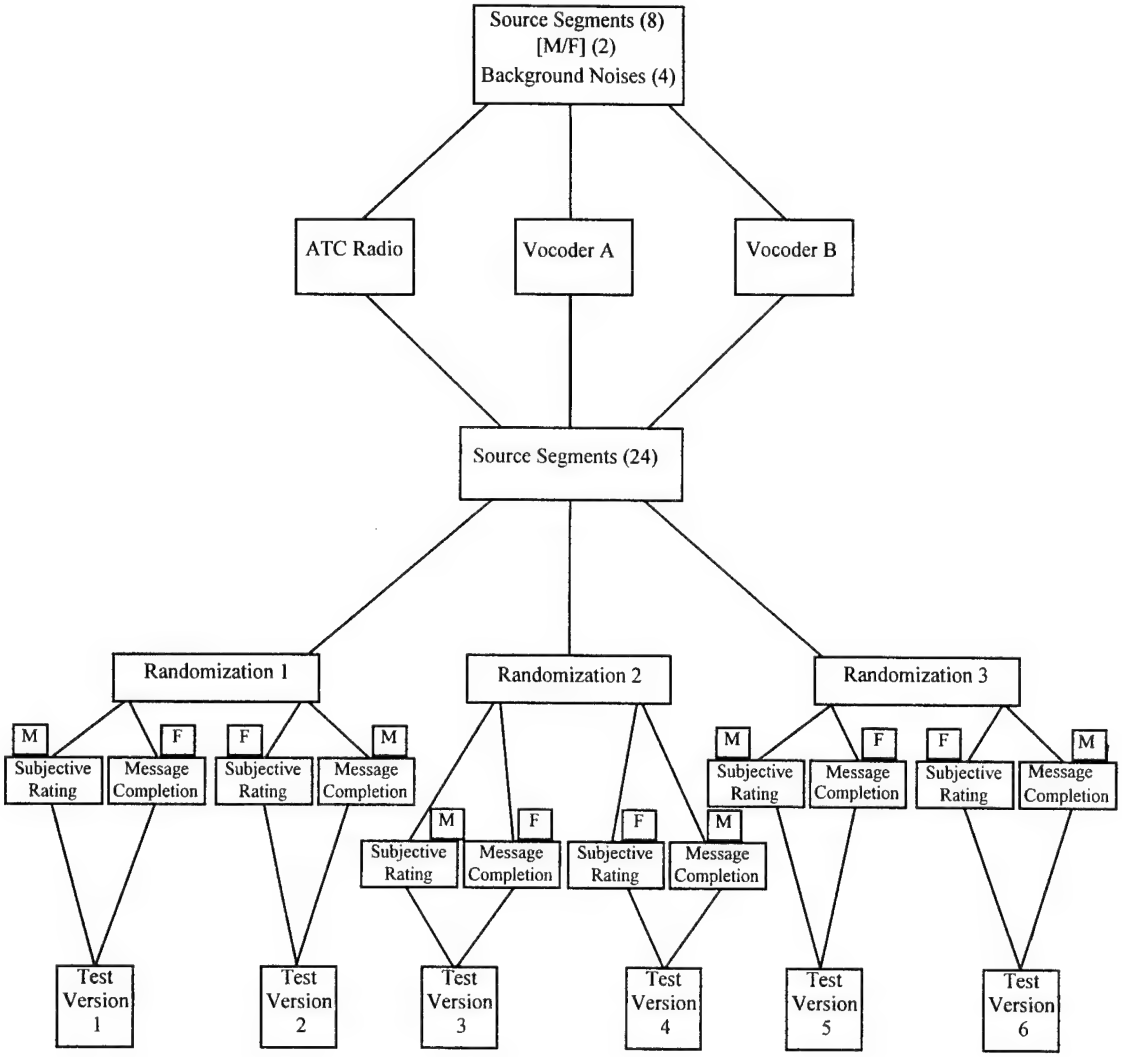

Figure 2. Field tape test construction from source tapes for subjective ratings and message completion tests.

tape was processed through each of the two vocoder models. The researchers presented eight of these messages to the participants in a random order during the evaluation.

Participants in the field listened through eight sets of Sony MDR V600 stereo headphones. These provided high audio quality and offered an enclosed headset that would reduce the influence of any noise in the testing room. The use of this type of headset was vital in the harsh noise environment of an ATCT or TRACON.

As many as eight participants at one time took part in the evaluation. Two stimulus tapes could be presented simultaneously to groups of four participants, each of the participants hearing one or the other of the tapes. The simultaneous presentation of two tapes allowed adjacently seated participants to hear different material to prevent cross-contamination from participant to

participant. A custom-made switch panel allowed headset switching to either stimulus tape and interconnected the stereo channels for presentation in both ears of the stereo headsets.

The master tapes played through two Sony TCD-D8 recorders. These were high-quality digital tape recorders. A Radio Shack 31-1991 stereo amplifier powered the headphone distribution panel. A fixed attenuator between the digital audio tape recorders and the amplifier allowed the amplifier to be set to a reproducible sound level for the duration of the study.

The headphone distribution panel provided outputs for two operator headphones to allow the operators to hear the material from the two source tapes. The distribution panel allowed for headset switching from channel to channel. The ability to switch channels replaced the need to plug and unplug headphones from the panel and allowed each of the two test operators to listen to each tape as necessary. The operator headsets allowed monitoring for the starting and stopping of the tape player as appropriate to pace the experiment to the participants' needs.

When participants arrived to the testing room, the researchers briefed the group about the purpose of the study, how it was to be conducted, and what was expected from them as participants. Each participant received a test booklet including a questionnaire requesting background information, including age, sex, experience, and current ATC status (see Appendix B). The researchers guaranteed the anonymity of the participants.

Eight or fewer participants were positioned around a testing table to minimize their interaction with each other. Initially, participants seated adjacently were to receive different audio stimuli to prevent cross contamination although this appeared unnecessary after a few sessions.

The two researchers played the audio source tapes to the participants. Each phrase was presented individually, pausing the tape to give time to write responses. The participants looked up at the operator when they were ready to receive the next voice message. When all participants were ready, the test monitors played the next stimulus phrase from the tape.

2.3.4 Sample Size and Classification

To insure a representative sample of participants, the data acquisition occurred from visits to two ARTCCs and four ATCT/TRACONs. Site selection was random within each region chosen. Visits to ARTCCs were 3 to 4 days in duration. ATCT/TRACON testing lasted 3 days. In total, 207 controllers participated with 90 working in an ARTCC environment and the remaining in an ATCT or TRACON environment. Table 5 gives a summary of the field test locations.

The background questionnaire yielded the following information concerning the participants:

a. The sample consisted of 87% male controllers and 13% female controllers.

b. The average age in years of the controllers tested was 38.5 years with a standard deviation of 6.1.

c. The average years of experience controlling traffic (including any previous military experience) was 14.1 years with a standard deviation of 6.7.

Table 5. Field Testing Air Traffic Facilities

Facility Description	*Region*	*Number of Participants*
ATCT/TRACON	Great Lakes	31
ATCT/TRACON	Southwest	26
ATCT/TRACON	Western Pacific (1)	35
ATCT/TRACON	Western Pacific (2)	25
ARTCC	Southern	44
ARTCC	Northwest Mountain	46

d. The average months in the past year that the participants had controlled traffic was 11.1 with a standard deviation of 2.9.

e. The status of the participants was 8% Developmental, 81% FPL, and 11% staff or supervisory personnel. For those participants at the FPL level, the average time at that level was 10.0 years with a standard deviation of 5.9.

f. The working environment of the participants was 44% ARTCC, 26% TRACON, 17% ATCT, and 13% worked both in TRACON and ATCT environments.

3. Results

3.1 Subjective Ratings Test

Tables 6, 7, 8, and 9 present means for the Subjective Ratings Test. Generally, the means for the intelligibility ratings were higher than the acceptability ratings. This result indicated that a majority of controllers' responses indicated a difference between what they could understand (intelligibility) and what they found acceptable.

Changes in the levels of the independent variables affect these means. In the next section, an Analysis of Variance (ANOVA) is presented to determine which of the independent variable effects are significant.

3.1.1 Main Effects and Interactions

Mean intelligibility and acceptability scores were computed for each of the test conditions based upon the ratings of the three replications per condition. These scores were subjected to a 2 x 4 x 3 ANOVA to determine which of the independent variables significantly affected the results. The ANOVA ascertained the relevance of the independent variables and their interactions for each of the dependent measures in the Subjective Ratings Test. Separate analyses were conducted for intelligibility and acceptability ratings.

Tables 10 and 11 illustrate the results of the ANOVAs for the main effects and interactions. Results at the α=.05 level are significant. Three-way interactions yielded significance for both dependent measures. Interactions between independent variables indicate that the main effects are not directly additive and you have to interpret the results in smaller sections. Since strong interactions were found, it became necessary to evaluate simple main effects.

Table 6. Mean Intelligibility Ratings as a Function of Background Noise and Equipment for Male Speaker Messages

Background	*Equipment*		
	Vocoder A	Vocoder B	Analog Radio
Jet	5.99 (1.62)	6.95 (1.21)	6.94 (1.25)
Propeller	6.28 (1.61)	6.98 1.27)	7.00 (1.25)
Helicopter	6.63 (1.35)	7.07 (1.10)	7.29 (1.28)
None	6.21 (1.68)	6.76 (1.43)	7.38 (0.92)
SD given in parenthesis *n*=207			

Table 7. Mean Intelligibility Ratings as a Function of Background Noise and Equipment for Female Speaker Messages

Background	*Equipment*		
	Vocoder A	Vocoder B	Analog Radio
Jet	5.81 (1.57)	6.41 (1.25)	6.78 (1.09)
Propeller	6.40 (1.28)	6.84 (1.07)	6.85 (1.14)
Helicopter	6.21 (1.38)	6.62 (1.20)	6.99 (0.96)
None	7.19 (1.01)	7.24 (1.08)	7.31 (0.81)
SD given in parenthesis *n*=207			

Table 8. Mean Acceptability Ratings as a Function of Background Noise and Equipment for Male Speaker Messages

Background	*Equipment*		
	Vocoder A	Vocoder B	Analog Radio
Jet	3.34 (1.43)	4.41 (1.64)	3.24 (1.68)
Propeller	3.88 (1.55)	4.73 (1.68)	3.21 (1.66)
Helicopter	4.37 (1.65)	4.59 (1.65)	4.13 (1.80)
None	4.56 (2.07)	5.35 (1.92)	4.35 (1.75)
SD given in parenthesis *n*=207			

Table 9. Mean Acceptability Ratings as a Function of Background Noise and Equipment for Female Speaker Messages

Background	*Equipment*		
	Vocoder A	Vocoder B	Analog Radio
Jet	3.08 (1.47)	3.84 (1.52)	3.36 (1.59)
Propeller	4.00 (1.52)	4.60 (1.60)	3.21 (1.49)
Helicopter	3.41 (1.55)	3.36 (1.61)	3.36 (1.69)
None	6.53 (1.57)	6.46 (1.77)	3.85 (1.73)
SD given in parenthesis *n* = 207			

Table 10. ANOVA for Intelligibility Ratings

Source of Variation	*Degrees of Freedom*	*Sum of Squares*	*Mean Square*	*F Statistic*	*Result*
Sex (S)	1	8.56	8.56	0.31	Not Significant
Background (B)	3	267.17	89.06	77.57	Significant*
Equipment (E)	2	684.98	342.49	123.99	Significant*
S x B	3	202.05	67.35	58.67	Significant*
S x E	2	34.03	17.02	6.16	Significant*
B x E	6	56.58	9.43	11.40	Significant*
S x B x E	6	72.78	12.13	14.67	Significant*
$*p < .05$					

Table 11. ANOVA for Acceptability Ratings

Source of Variation	*Degrees of Freedom*	*Sum of Squares*	*Mean Square*	*F Statistic*	*Result*
Sex (S)	1	15.30	15.30	0.37	Not Significant
Background (B)	3	2854.29	951.43	331.05	Significant*
Equipment (E)	2	1417.97	708.99	100.73	Significant*
S x B	3	791.83	263.94	91.84	Significant*
S x E	2	90.12	45.06	6.40	Significant*
B x E	6	668.73	111.45	65.39	Significant*
S x B x E	6	450.94	75.16	44.09	Significant*
$*p < .05$					

3.1.2 Simple Main Effects

To determine the simple main effect of equipment on the intelligibility rating, a simple main effects analysis was performed by varying the independent variables of sex and background noise through their levels. Table 12 shows the results of this analysis.

Except for the combination of no background noise and a female speaker, the simple main effect of equipment was significant for the intelligibility rating. A significant simple main effect means that, for example, a male speaker with jet background noise exhibits differences in the mean intelligibility ratings for the two vocoders and analog radio. Since these significant simple main effects were found, post hoc Tukey Honestly Significant Difference (HSD) tests were conducted to determine which factor levels were significantly different from the others. The results are presented both in tabular form to indicate which means are significantly different and in graphical form to allow the reader to visualize the interactions of the test conditions.

Table 12. Analysis of Simple Main Effects for Equipment Intelligibility Ratings

Sex	*Background*	*F Statistic*	*Result*
Male	Jet	64.16	Significant*
Male	Propeller	28.64	Significant*
Male	Helicopter	25.10	Significant*
Male	None	63.00	Significant*
Female	Jet	51.47	Significant*
Female	Propeller	23.06	Significant*
Female	Helicopter	54.96	Significant*
Female	None	1.06	Not Significant
$*p < .05$			

Table 13 illustrates the results of the Tukey tests with the means listed from highest to lowest values. The means connected by bars are not significantly different and those not connected by bars are significantly different.

Except for cases of a female speaker and no background noise, radio and vocoder B are always more intelligible than vocoder A. Radio is rated more intelligible than vocoder B except in the cases of a female speaker and no background noise, a male speaker and jet background noise, and a male speaker and propeller background noise. Figures 3 and 4 represent these means graphically.

A simple main effects analysis of equipment on the acceptability rating yielded the results presented in Table 14. Except for the condition with helicopter background noise and female speaker, the simple main effect of equipment on the acceptability rating was significant. A post hoc Tukey HSD test determined which means differed associated to equipment. The results of the Tukey test performed are in Table 15 where the same conventions apply as previously seen in Table 13.

Table 15 reveals that vocoder B is preferable on the acceptability scale to both vocoder A and normal radio. In addition, the test participants sampled preferred vocoder A to analog radio on the acceptability scale although, in a majority of the cases studied, the means did not differ. In all cases vocoder B scored significantly higher than radio except for the conditions of female speaker and helicopter background noise in which no simple main effect existed. Figures 5 and 6 depict a graphical representation of the acceptability means.

Table 13. Tukey HSD Post Hoc Comparisons for Equipment Intelligibility Ratings

Sex of Speaker	*Background*	*Result*	*Interpretation*
Male	Jet	$\overline{BR}A$	Vocoder B and Radio means do not differ. Both means are higher than Vocoder A mean.
Male	Propeller	$\overline{BR}A$	Vocoder B and Radio means do not differ. Both means are higher than Vocoder A mean.
Male	Helicopter	RBA	All means are different. Radio mean is higher than Vocoder B mean, which is higher than Vocoder A mean.
Male	None	RBA	All means are different. Radio mean is higher than Vocoder B mean, which is higher than Vocoder A mean.
Female	Jet	RBA	All means are different. Radio mean is higher than Vocoder B mean, which is higher than Vocoder A mean.
Female	Propeller	$\overline{BR}A$	Vocoder B and Radio means do not differ. Both means are higher than Vocoder A mean.
Female	Helicopter	RBA	All means are different. Radio mean is higher than Vocoder B mean, which is higher than Vocoder A mean.
Female	None	NS	Not Significant

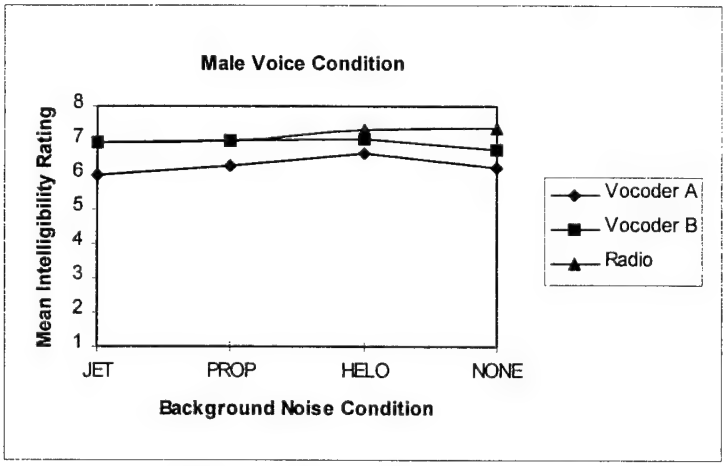

Figure 3. Mean intelligibility ratings as a function of equipment and background noise for male speaker messages.

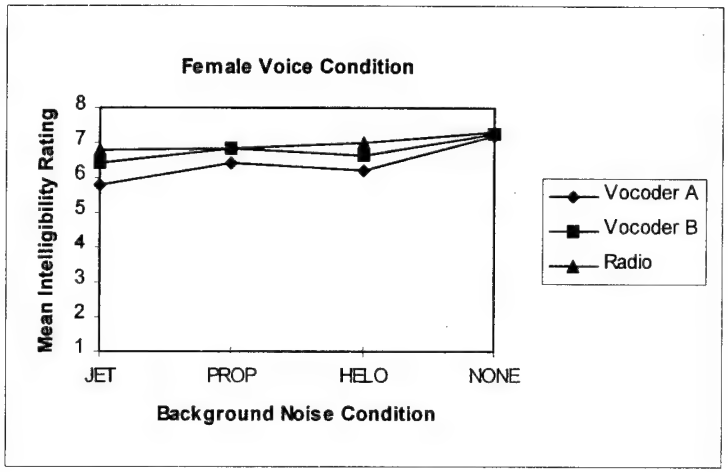

Figure 4. Mean intelligibility ratings as a function of equipment and background noise for female speaker messages.

Table 14. Analysis of Simple Main Effects for Equipment Acceptability Ratings

Sex of Speaker	*Background*	*F Statistic*	*Result*
Male	Jet	54.15	Significant*
Male	Propeller	63.32	Significant*
Male	Helicopter	4.93	Significant*
Male	None	16.56	Significant*
Female	Jet	24.03	Significant*
Female	Propeller	59.33	Significant*
Female	Helicopter	0.14	Not Significant
Female	None	149.47	Significant*

*$p < .05$

Table 15. Tukey HSD Post Hoc Comparisons for Equipment Intelligibility Ratings

Sex of Speaker	*Background*	*Result*	*Interpretation*
Male	Jet	$B\overline{AR}$	Vocoder B is higher and different from Vocoder A and Radio means, which are the same.
Male	Propeller	BAR	All means are different. Vocoder B mean is higher than Vocoder A mean, which is higher than Radio mean.
Male	Helicopter	$\overline{BAR}$	Vocoder B and Vocoder A means are the same. Vocoder A and Radio means are the same. Vocoder B mean is higher and different than Radio mean.
Male	None	$B\overline{AR}$	Vocoder B mean is higher and different than Vocoder A and Radio means, which are the same.
Female	Jet	$B\overline{RA}$	Vocoder B mean is higher and different than Vocoder A and Radio means, which are the same.
Female	Propeller	BAR	All means are different. Vocoder B mean is higher than Vocoder A mean, which is higher than Radio mean.
Female	Helicopter	NS	Not Significant
Female	None	$\overline{ABR}$	Vocoder A and Vocoder B means are the same and higher than Radio mean, which is different.

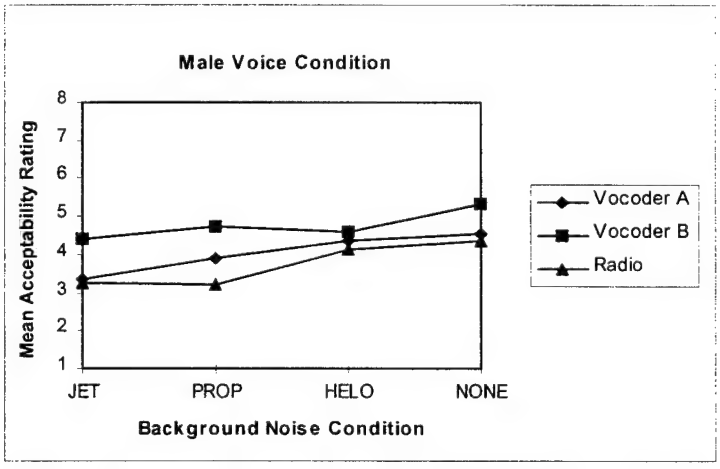

Figure 5. Mean acceptability ratings as a function of equipment and background noise for male speaker messages.

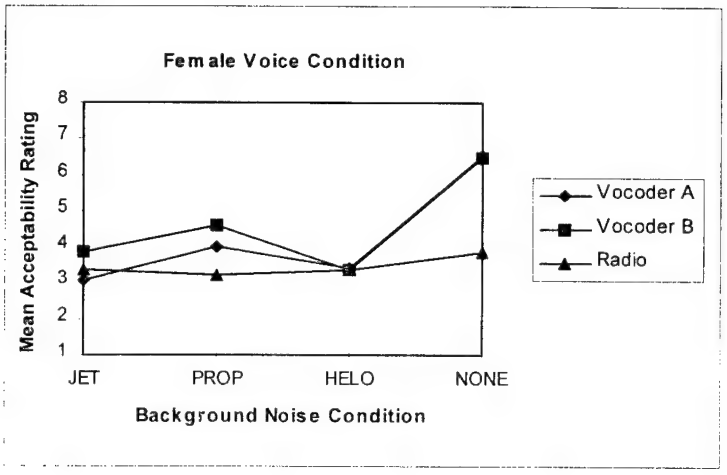

Figure 6. Mean acceptability ratings as a function of equipment and background noise for female speaker messages.

To determine the simple main effect of background noise on intelligibility rating, an analysis of simple main effects was performed by varying the independent variables of sex and equipment noise through their levels. Table 16 exhibits the results of this analysis.

Table 16. Analysis of Simple Main Effects for Background Noise Intelligibility Ratings

Sex of Speaker	*Equipment*	*F Statistic*	*Result*
Male	Vocoder A	22.32	Significant*
Male	Vocoder B	4.80	Significant*
Male	Radio	17.45	Significant*
Female	Vocoder A	75.80	Significant*
Female	Vocoder B	44.09	Significant*
Female	Radio	32.75	Significant*
$*p < .05$			

In all cases, the simple main effect of background noise was significant. A post hoc Tukey HSD test determined which means differed. The results of the Tukey test are presented in Table 17. The results are complex, but some generalities arise. Cases with no background noise perform well as expected and tend to separate from the field, especially in cases where the speaker was female. In addition, helicopter noise exhibited high averages although it did not tend to divide itself from the field. Jet background scored very poorly for the intelligibility rating especially when the speaker was female and the transmission was through vocoder A. Figures 7 and 8 depict these means graphically.

Table 17. Tukey HSD Post Hoc Comparisons for Background Noise Intelligibility Ratings

Sex of Speaker	*Equipment*	*Result*	*Interpretation*
Male	Vocoder A	$H\overline{PNJ}$	Helicopter background mean is higher and is different than propeller, no background, and jet background, which are the same.
Male	Vocoder B	$\overline{HPJN}$	Helicopter mean is higher than and different from no background mean. All other means are the same.
Male	Radio	$\overline{NH}\,\overline{PJ}$	No background and helicopter background means are the same and are higher than propeller background and jet background means, which are the same.
Female	Vocoder A	$N\overline{PHJ}$	No background is higher than propeller and helicopter backgrounds, which are the same and higher than jet background, which is different.
Female	Vocoder B	$N\overline{PHJ}$	No background is higher than propeller and helicopter backgrounds, which are the same. Propeller background higher and different than jet background.
Female	Radio	$N\overline{HPJ}$	No background is higher than helicopter, propeller, and jet backgrounds, which are the same.

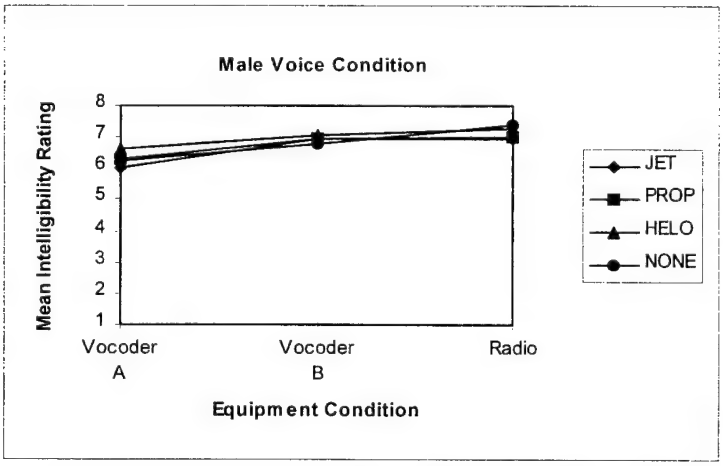

Figure 7. Mean intelligibility ratings as a function of background noise and equipment for male speaker messages.

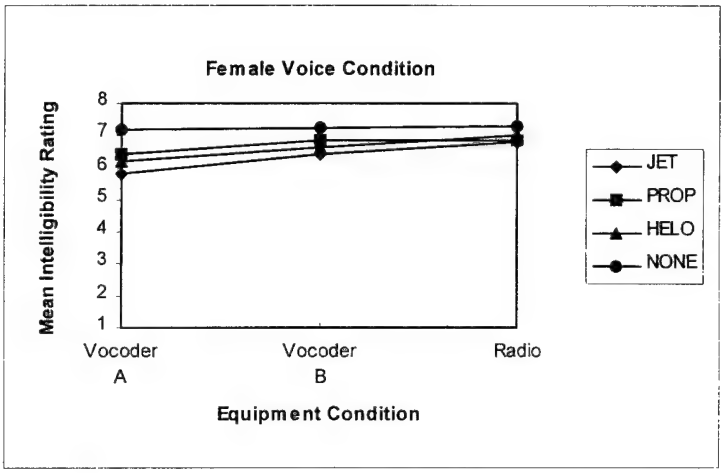

Figure 8. Mean intelligibility ratings as a function of background noise and equipment for female speaker messages.

To determine the simple main effect of background noise on the acceptability rating, an analysis of simple main effects was performed on this independent variable. The results are presented in Table 18.

Table 18. Analysis of Simple Main Effects for Background Noise Acceptability Ratings

Sex	*Equipment*	*F Statistic*	*Result*
Male	Vocoder A	42.46	Significant*
Male	Vocoder B	16.85	Significant*
Male	Radio	80.45	Significant*
Female	Vocoder A	332.42	Significant*
Female	Vocoder B	207.10	Significant*
Female	Radio	22.51	Significant*
$*p < .05$			

In all cases, the simple main effect of background noise on the acceptability rating was significant. A post hoc Tukey test determined which means differed. Table 19 exhibits the results of the Tukey test performed.

Table 19. Tukey HSD Post Hoc Comparisons for Background Noise Acceptability Ratings

Sex of Speaker	*Equipment*	*Result*	*Interpretation*
Male	Vocoder A	$\overline{NHPJ}$	No background and helicopter background means are the same and are higher than propeller background and jet background means, which are different.
Male	Vocoder B	$N\overline{PHJ}$	No background mean is higher than helicopter, propeller, and jet background means, which are the same.
Male	Radio	$\overline{NHPJ}$	No background and helicopter background means are the same and are higher than propeller backgrounds and jet background means, which are the same.
Female	Vocoder A	$NPHJ$	All means are different. No background mean is higher than propeller mean, which is higher than helicopter mean, which is higher than jet mean.
Female	Vocoder B	$NPJH$	All means are different. No background mean is higher than propeller background mean, which is higher than jet background mean, which is higher than helicopter background mean.
Female	Radio	$N\overline{PJH}$	No background mean is higher than helicopter, propeller, and jet background means, which are the same.

Mean acceptability rating scores under conditions of no background noise were higher and usually distinct from other means. In cases where the speaker was male, the helicopter background noise cases exhibited high marks in acceptability relative to the other background noises. However, this trend did not hold when the speaker was female. Jet background noise showed low scores in all cases as in the intelligibility rating scores. Figures 9 and 10 depict the means corresponding to this analysis graphically. The graphs suggest that vocoder B is in no case worse than analog radio and, with some backgrounds, is decidedly better than analog radio.

Analysis of the simple main effect of sex of speaker yielded useful information concerning how the pitch and tone of the speaker interacted with the other independent variables. Table 20 shows the results. A noteworthy feature of the simple main effect of sex is the frequency of significant effects exhibited for the vocoders versus that of the analog radio control. Since there are only two categories associated with this independent variable, post hoc Tukey tests were not a requirement. The summary of the results is presented in Table 21.

Across all conditions, there appears to be no preference for either sex in intelligibility ratings. Each case seems to behave differently with changes in equipment and background noise. Figures 11, 12 and 13 depict these means graphically. The similarity in the shape of the performance curves for vocoders A and B is noteworthy.

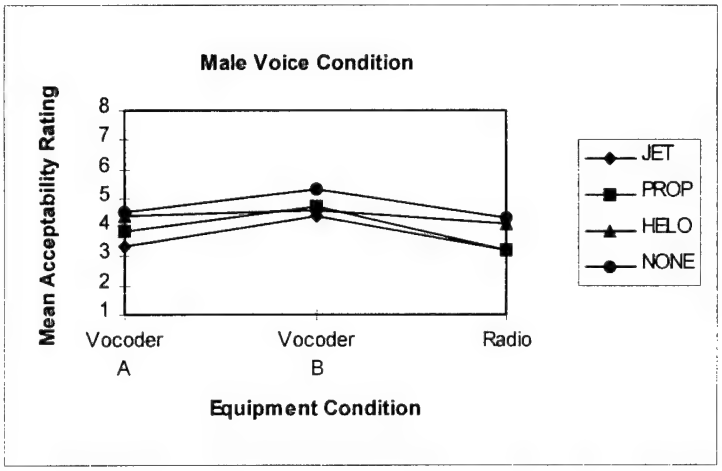

Figure 9. Mean acceptability ratings as a function of background noise and equipment for male speaker messages.

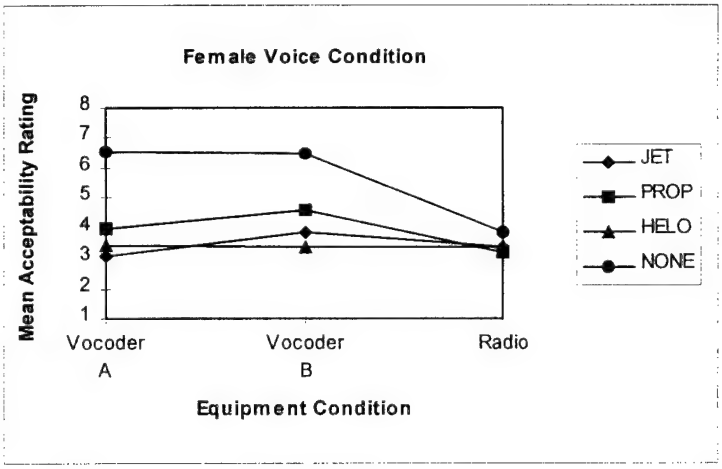

Figure 10. Mean acceptability ratings as a function of background noise and equipment for female speaker messages.

Table 20. Analysis of Simple Main Effects for Speaker Intelligibility Ratings

Equipment	*Background*	*F Statistic*	*Result*
Vocoder A	Jet	0.88	Not Significant
Vocoder A	Propeller	0.44	Not Significant
Vocoder A	Helicopter	6.38	Significant*
Vocoder A	None	35.50	Significant*
Vocoder B	Jet	14.27	Significant*
Vocoder B	Propeller	1.12	Not Significant
Vocoder B	Helicopter	10.46	Significant*
Vocoder B	None	11.12	Significant*
Radio	Jet	1.24	Not Significant
Radio	Propeller	0.98	Not Significant
Radio	Helicopter	5.79	Significant*
Radio	None	0.39	Not Significant
* $p < .05$			

Table 21. Tukey HSD Post Hoc Comparisons for Speaker Intelligibility Ratings

Equipment	*Background*	*Interpretation*
Vocoder A	Jet	Not Significant
Vocoder A	Propeller	Not Significant
Vocoder A	Helicopter	Male voice mean intelligibility rating is higher than female voice mean intelligibility rating
Vocoder A	None	Female voice mean intelligibility rating is higher than male voice mean intelligibility rating
Vocoder B	Jet	Male voice mean intelligibility rating is higher than female voice mean intelligibility rating
Vocoder B	Propeller	Not Significant
Vocoder B	Helicopter	Male voice mean intelligibility rating is higher than female voice mean intelligibility rating
Vocoder B	None	Female voice mean intelligibility rating is higher than male voice mean intelligibility rating
Radio	Jet	Not Significant
Radio	Propeller	Not Significant
Radio	Helicopter	Male voice mean intelligibility rating is higher than female voice mean intelligibility rating
Radio	None	Not Significant

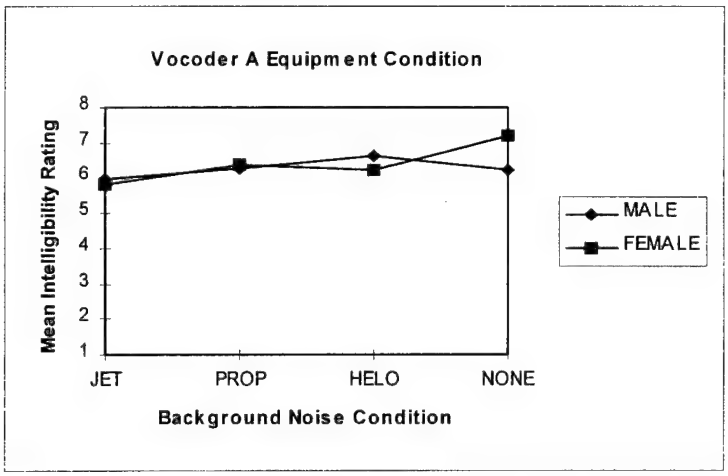

Figure 11. Mean intelligibility ratings as a function of sex of speaker and background noise for vocoder A messages.

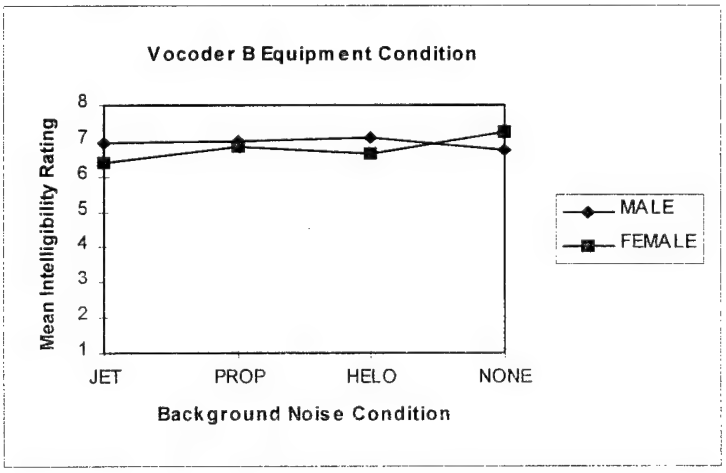

Figure 12. Mean intelligibility ratings as a function of sex of speaker and background noise for vocoder B messages.

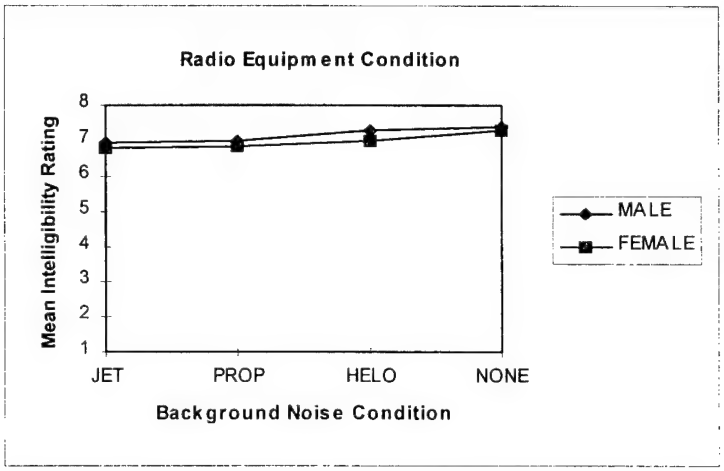

Figure 13. Mean intelligibility ratings as a function of sex of speaker and background noise for analog radio messages.

Table 22 depicts the simple main effect of sex of speaker on the acceptability rating dependent measure. An important characteristic of the simple main effect of sex on the acceptability rating is that the significant effects occur with the same combinations of equipment and background as the intelligibility rating analysis. Moreover, these effects occur with identical ordering in the means as the intelligibility rating. This is exhibited in the results for the acceptability rating in Table 23.

Table 22. Analysis of Simple Main Effects for Speaker Acceptability Ratings

Equipment	*Background*	*F Statistic*	*Result*
Vocoder A	Jet	2.24	Not Significant
Vocoder A	Propeller	0.51	Not Significant
Vocoder A	Helicopter	24.94	Significant*
Vocoder A	None	79.90	Significant*
Vocoder B	Jet	8.83	Significant*
Vocoder B	Propeller	0.45	Not Significant
Vocoder B	Helicopter	38.44	Significant*
Vocoder B	None	26.05	Significant*
Radio	Jet	0.36	Not Significant
Radio	Propeller	0.00	Not Significant
Radio	Helicopter	12.12	Significant*
Radio	None	5.39	Significant*

* $p < .05$

Table 23. Tukey HSD Post Hoc Comparisons for Speaker Acceptability Ratings

Equipment	*Background*	*Interpretation*
Vocoder A	Jet	Not Significant
Vocoder A	Propeller	Not Significant
Vocoder A	Helicopter	Male voice mean acceptability rating is higher than female voice mean acceptability rating
Vocoder A	None	Female voice mean acceptability rating is higher than male voice mean acceptability rating
Vocoder B	Jet	Male voice mean acceptability rating is higher than female voice mean acceptability rating
Vocoder B	Propeller	Not Significant
Vocoder B	Helicopter	Male voice mean acceptability rating is higher than female voice mean acceptability rating
Vocoder B	None	Female voice mean acceptability rating is higher than male voice mean acceptability rating
Radio	Jet	Not Significant
Radio	Propeller	Not Significant
Radio	Helicopter	Male voice mean intelligibility rating is higher than female voice mean intelligibility rating
Radio	None	Male voice mean intelligibility rating is higher than female voice mean intelligibility rating

Across all conditions, there appears to be no preference for either sex in acceptability ratings. The means for these cases are depicted graphically in Figures 14, 15, and 16. The patterns are similar to the intelligibility ratings for this simple main effect, which could indicate some correlation between intelligibility and acceptability ratings. Such a correlation will be investigated in the next section.

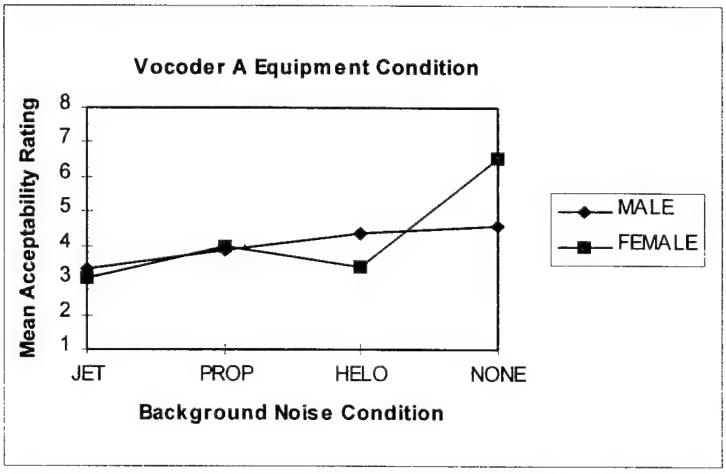

Figure 14. Mean acceptability ratings as a function of sex of speaker and background noise for vocoder A messages.

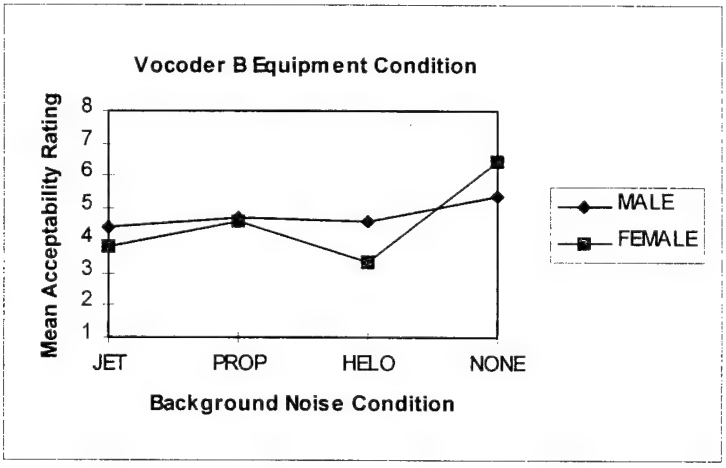

Figure 15. Mean acceptability ratings as a function of sex of speaker and background noise for vocoder B messages.

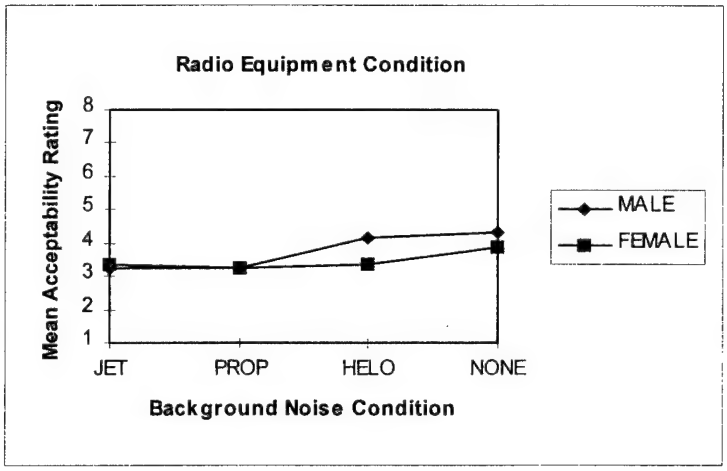

Figure 16. Mean acceptability ratings as a function of sex of speaker and background noise for analog radio messages.

3.1.3 Subjective Ratings Correlational Analysis

To ascertain the relationship between the participants' intelligibility ratings and acceptability ratings, a Pearson product moment coefficient of correlation was calculated between the intelligibility rating and the acceptability rating. The correlation yielded a result of .37 with 7,344 observations. Further analysis indicated the effects of a correlation between the two dependent variables as a function of equipment. The results are presented in Table 24.

Table 24. Correlations Between Intelligibility and Acceptability Ratings by Equipment

Equipment	*Observations*	*Pearson Product Moment Coefficient*
Vocoder A	2,448	.42
Vocoder B	2,448	.44
Radio	2,448	.37
n=207		

The results are in the .4 range indicating a positive linear relationship between that which the participants found intelligible and that which was acceptable. Given the low correlations, however, the reader may conclude relative independence of ratings between intelligibility and acceptability.

3.2 Message Completion Test

The results of the Message Completion Test indicated high scores in this objective measure of intelligibility across all categories. Tables 25 and 26 show a listing of these means under each of

Table 25. Mean Message Completion Test Scores as a Function of Background Noise and Equipment for Male Speaker Messages

Background	*Equipment*		
	Vocoder A	Vocoder B	Analog Radio
Jet	8.79 (.50)	8.90 (.31)	8.80 (.43)
Propeller	8.86 (.40)	8.93 (.27)	8.95 (.26)
Helicopter	8.92 (.30)	8.93 (.35)	8.94 (.34)
None	8.81 (.52)	8.94 (.23)	8.89 (.45)
SD given in parenthesis *n*=207			

Table 26. Mean Message Completion Test Scores as a Function of Background Noise and Equipment for Female Speaker Messages

Background	*Equipment*		
	Vocoder A	Vocoder B	Analog Radio
Jet	8.79 (.59)	8.81 (.53)	8.82 (.41)
Propeller	8.98 (.14)	8.95 (.26)	8.93 (.29)
Helicopter	8.88 (.35)	8.84 (.50)	8.90 (.32)
None	8.93 (.25)	8.93 (.25)	8.93 (.25)
SD given in parenthesis *n*=207			

the test conditions. As previously noted, each participant evaluated three messages per test condition. Each message contained three blanks to complete leading to a possible perfect score of 9 correct responses per participant per test condition. As is indicated in the tables, the mean scores are very near that of perfect responses. An ANOVA ascertained if significant differences existed for these means. This analysis was similar to the analysis completed for the Subjective Ratings Test and the results are presented in Table 27.

Table 27. ANOVA for Message Completion Test Scores

Source of Variation	*Degrees of Freedom*	*Sum of Squares*	*Mean Square*	*F Statistic*	*Result*
Sex (S)	1	.00	.00	0.03	Not Significant
Background (B)	3	1.69	.56	14.52	Significant*
Equipment (E)	2	.19	.10	2.20	Not Significant
S x B	3	.38	.13	3.23	Significant*
S x E	2	.30	.15	3.47	Significant*
B x E	6	.20	.03	0.74	Not Significant
S x B x E	6	.20	.03	0.74	Not Significant
$*p < .05$					

The ANOVA yielded two-way interactions of sex and background and sex and equipment. An analysis of the two-way interactions ascertained the significant combinations of the interacting variables.

ANOVAs conducted to determine the interaction of sex of speaker and equipment revealed a significant effect when the speaker was male. Post hoc Tukey testing determined that vocoder B was objectively superior in understandability to vocoder A under the male voice condition. However, the testing revealed no conclusions regarding the comparison of vocoders to radio. Figure 17 depicts a graphical representation of the mean intelligibility scores associated with this analysis.

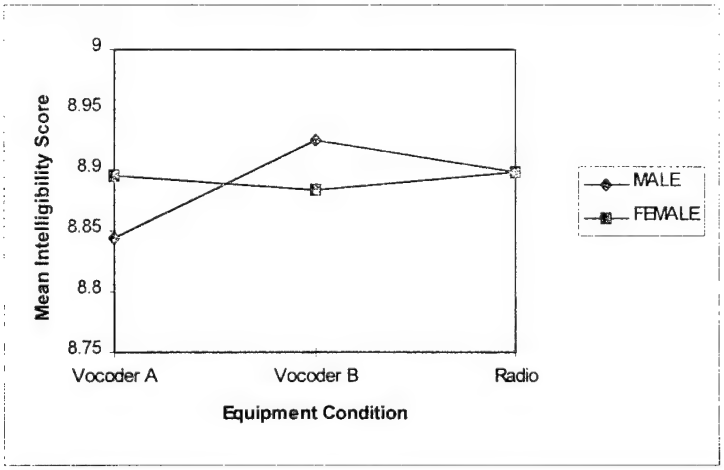

Figure 17. Mean scores from the Message Completion Test as a function of sex of speaker and equipment.

ANOVA testing for interaction of sex with background noise revealed that for both conditions of the speaker, the effect of background noise was significant. Post hoc Tukey tests determined which means differed. Table 28 depicts the results of the Tukey HSD tests for the interaction of sex of speaker and background noise. Cases with jet background noise appear to score low in this objective measure of intelligibility. This result is in agreement with the subjective ratings of intelligibility previously presented. Figure 18 shows a graphical representation of these means.

Table 28. Tukey HSD Post Hoc Background Noise Comparisons for Message Completion Test Scores

Sex of Speaker	*Result*	*Interpretation*
Male	$\overline{\overline{HPNJ}}$	Helicopter, propeller, and no background noise means are the same. No background and jet means are the same. Jet background mean is lower and different than helicopter and propeller means.
Female	$\overline{\overline{PNHJ}}$	Propeller background and no background means are the same. No background and helicopter background means are the same. Helicopter mean is lower and different than propeller mean. Jet background mean is lower and different than all other means.

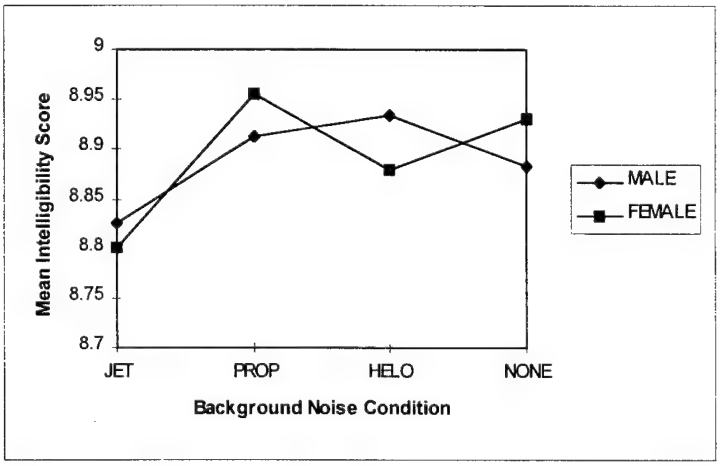

Figure 18. Mean scores from the Message Completion Test as a function of sex of speaker and background noise.

3.3 Audio Preference Test

The data from this test were analyzed by counting the number of participants who preferred vocoder model A and vocoder model B for each test condition. Individual chi-squared analysis determined the effects of each experimental condition on vocoder preference.

3.3.1 Proportional Responses

Of the 207 controllers who participated in the field testing, 199 completed the Audio Preference Test successfully. Those who did not failed to make a choice on all eight responses or chose both vocoders for at least one response. Table 29 illustrates the responses by test condition. Across all test conditions, 86% of participants selected vocoder B. A larger majority (89%) preferred Vocoder B when the speaker was male than when the speaker was female (83%). The effect of the background noise is also exhibited. The largest preference for vocoder B occurred with the propeller background noise condition (91%) with the lowest preference under the helicopter background noise condition (80%).

Table 29 . Response Frequencies from the Audio Preference Test as a Function of Background Noise and Sex of Speaker

Background Noise	*Vocoder A % (Freq)*	*Vocoder B % (Freq)*	*Sex of Speaker*	*Vocoder A % (Freq)*	*Vocoder B % (Freq)*
Jet	10 % (38)	90 % (360)	Male	11 % (85)	89 % (711)
Propeller	9 % (35)	91 % (363)	Female	17 % (139)	83 % (657)
Helicopter	20 % (78)	80 % (320)	Total	14 % (224)	86 % (1368)
None	18 % (73)	82 % (325)			
Total	14 % (224)	86 % (1368)			
n=207					

3.3.2 Chi-Square Analysis

To determine if a statistically viable preference occurred for one vocoder model, a chi-square analysis was performed on the results of the Audio Preference Test. Table 30 illustrates the results for this test. For all conditions listed, the calculated value of the chi-square test statistic far exceeded the critical value for the rejection of the null hypothesis. Therefore, there was a strong preference for Vocoder B. The high values for the calculated chi-square test statistics reflected the large sample size and strong deviation from the no-preference expectation.

Table 30. Chi-Square Analysis for the Audio Preference Test

Sex of Speaker	*Background Noise*	*Vocoder A*	*Vocoder B*	*Chi-Square Statistic*	*Result*
Male	Jet	13	186	150.40	Significant*
Male	Propeller	15	184	143.52	Significant*
Male	Helicopter	21	178	123.86	Significant*
Male	None	36	163	81.05	Significant*
Female	Jet	25	174	111.56	Significant*
Female	Propeller	20	179	127.04	Significant*
Female	Helicopter	57	142	36.31	Significant*
Female	None	37	162	78.52	Significant*

$n = 199$
$*p < .05$

3.3.3 Preference Rationale

As previously noted, the test participants were requested to give a rationale for their preference of vocoder for each test condition. To analyze these preferences, a taxonomy was created for the classification of responses.

Not all participants responded to the query. There were no instructions given to the participants concerning the structure of their responses. Therefore, the focus of the responses varied. Further, each participant's preference did not necessarily state reasons directly supportive of their choice. Rather, the response may have included reasons for not selecting the message presented from the other vocoder or, by direct comparison, the two vocoders. If a participant's response referred to the contrasting vocoder, the response was transposed. For example, if a participant chose vocoder B and stated that vocoder A had too much background noise as the justification, it was presumed that vocoder B had less background noise and was classified accordingly.

Participants also responded with words like barreling, tinny, warble, and garble. Experts from the aviation field clarified the meaning of these terms to correctly include them in the taxonomy.

Many test participants responded with multiple reasons for their choice. When multiple reasons were given, distinct reasons (those included in different classifications) were parsed. The number of responses given were then placed with the participant's identification number which would enable weighting the responses if the need arose.

Table 31 gives the classification used to categorize the responses. Thirteen taxonomy codes fit all the responses, the most general of which included all non-repeated responses that were deemed to fit no category. These codes were empirically derived from the controller responses rather than generated a priori.

Table 31. Classification of Audio Preference Responses

Taxonomy Code	*Classification*
1	natural/realistic/human-like
2	clear, better intelligibility (easier to understand), better voice quality, smooth/less modulation, less warbled
3	less background noise (less static or scratchy, less echo/reverberation/barreling)
4	excessive bass in voice
5	less garbled, less distortion/less muffled
6	less broken, choppy or clipped
7	less distracting
8	loudness
9	less "tinny" and/or "tunnely"
10	sound tone or pitch
11	no difference, guess
12	both poor/did not like either
13	fits no category

As previously noted, there were 199 participants who successfully completed the Audio Preference Test. Each participant had eight opportunities to assert their preference of vocoder, resulting in 1592 total possible responses. Of these, 329 instances arose where the participants gave no rationale for their response. Multiple responses for a single test condition from a single participant were included in the taxonomy as it was unclear which response held the strongest weight. The rationale was not used to determine vocoder preference. It gave justification for vocoder preference precluding the need for exacting statistics. As Figure 19 illustrates, the largest number of responses fell into categories two, three, and five reflecting preferred intelligibility effects in the voice and less background noise for the favored vocoder B. The small number of responses that were unclassifiable (taxon 13) indicates the comprehensiveness of the empirical taxonomy.

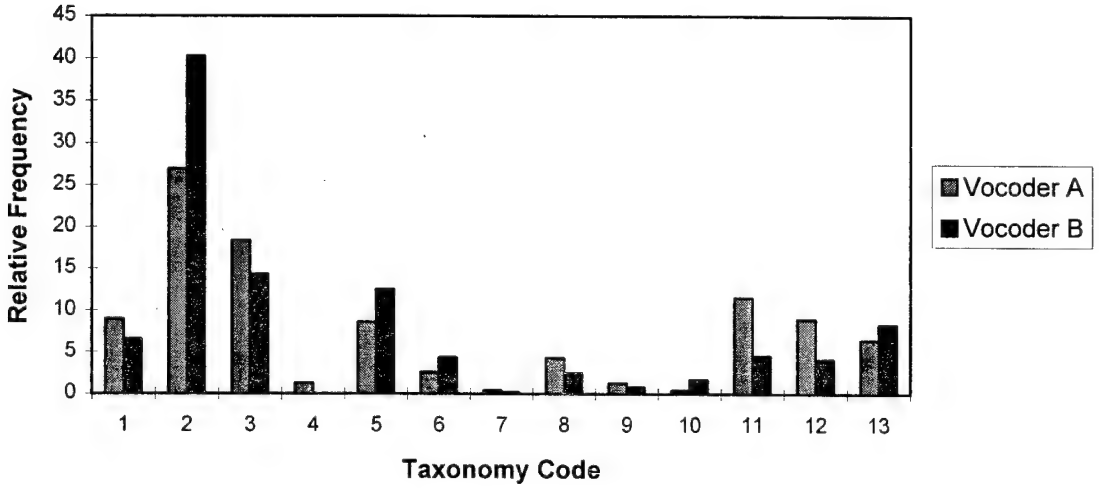

Figure 19. Relative frequencies across categories for audio preference test selection rationale.

3.4 Exit Survey

Table 32 shows the results of the exit survey. The mean and standard deviation are with respect to a 10-point scale with the anchors as shown. As illustrated, the participants were generally not knowledgeable in vocoder technology prior to the study, which is a favorable indication of an unbiased group. The enthusiasm level was somewhat high for the group as was the participants' value of the time spent performing the evaluation. These aspects add validity to the data obtained.

4. Discussion

4.1 Analysis of Subjective Ratings

Analysis of the subjective ratings for both intelligibility and acceptability revealed three-way interactions of the independent variables of equipment, background noise, and sex of speaker. Interactions are basically statistical results that mean that the impacts of the various independent variables do not add up neatly. This made the analysis more challenging but within capability. It is unlikely that one explanation could account for all the interactions. However, one possible source for the interactions could lie in the voice and background noise behaving as a system when processed through the speech compression algorithm of the vocoders rather than two independent sound sources. This nonlinear type of behavior could explain the interactions seen regarding the participant ratings of both intelligibility and acceptability of the vocoder processed messages. These interaction effects are less prominent when the equipment variable is normal analog radio, which was expected to behave linearly with respect to the input signal.

Table 32. Exit Survey Results

Query	*Scale Anchors*		*Mean*	*SD*
	(1)	*(10)*		
Your knowledge of vocoders prior to this study.	*not knowledgeable*	*extremely knowledgeable*	1.66	1.34
Your enthusiasm to participate in the study.	*not enthusiastic*	*extremely enthusiastic*	7.00	1.92
The comfort level of the headphones.	*not comfortable*	*extremely comfortable*	8.40	1.66
Your current state of health.	*not healthy*	*extremely healthy*	8.93	1.35
The professionalism of the presentation.	*not professional*	*extremely professional*	9.42	0.84
The value of the time spent on this session.	*not valuable*	*extremely valuable*	7.90	1.58
The realism of the aircraft background noises.	*not realistic*	*extremely realistic*	6.89	2.10

4.1.1 Effect of Equipment

The fact that controllers rated analog radio generally more intelligible than vocoders is not surprising. Analog radio was expected to perform better due to the simplifications made to model the speech in the vocoders. Of special note, however, is that in five of eight conditions analog radio was indistinguishable in intelligibility to vocoder B. The largest difference in the intelligibility means between analog radio and vocoder B was .62 and occurred with a male speaker and no background noise. This difference reflects only 9% of the effective range of the 8-point scale used.

Analysis of the simple main effect of equipment on the acceptability rating revealed that vocoder B was superior to analog radio for all significant combinations of sex of speaker and background noise. Further, under limited conditions where the tests revealed differences in the means between vocoder A and radio, vocoder A was also superior to radio on the acceptability scale. One possible cause for the high ratings of the vocoders could lie in the clarity of the digital signal that is void of the analog radio signal static. The largest differences in acceptability between the higher rated vocoder B and radio occurred with propeller background noise. Under conditions of male and female speakers, the differences in the means are 1.52 and 1.39, respectively. These numbers reflect approximately 20% of the effective range of the 8-point scale, reflecting a strong preference of vocoder B over analog radio. Overall, the controllers responded very favorably to speech processed by vocoder B.

4.1.2 Effect of Background Noise

The simple main effect of the background noise on the intelligibility rating measure revealed a consistent ordering of the means for the four background noise levels under the analog radio control. This trend was not apparent, however, with the vocoders tested. Cases with helicopter background noise rated higher in intelligibility than cases with no background noise for both vocoder A and B when the speaker was male. This trend reversed when the speaker was female where cases with no background noise received higher intelligibility scores than cases with helicopter background noise for both vocoders. The pattern of the ordering of the means with respect to background noise is generally consistent with both vocoders tested within each sex of speaker. Since the sound intensity level in decibels was the same for all the background noises in the study, the reader can conclude that sound quality aspects are responsible for the differences in the means between these cases. These aspects could include the manner in which the frequency spectrum of the background noise interferes with human speech, the manner in which the equipment processes the background noise and speech combination, or both effects.

Effects of the background noise on the acceptability rating revealed that cases with no background noise received generally higher scores for both analog radio and the vocoders tested. The ordering of the means was very similar for both vocoders and radio, which was not the case with the intelligibility rating. This is indicative that there were different characteristics of the voice message used to judge acceptability. Post hoc Tukey tests revealed that the most significant differences in the means for the background noise variable occurred with the vocoders when the speaker was female. This may be indicative of vocoder sensitivity to pitch and tone of the speaker. The next section discusses this effect more fully.

4.1.3 Effect of Sex of Speaker

The frequency of significant effects due to sex of speaker was higher with the vocoders tested than with the analog radio used as the control. This may be a result of the nonlinear effects described earlier. In all significant cases with background noise for both vocoders, the male speaker received higher ratings. In cases with no background noise for both vocoders, the female speaker received higher ratings. This may indicate intelligibility losses with the interaction of the female voice with the background noise as, without any background noise, it was the preferred voice on the intelligibility scale. It may also be the result of bias or preference on the part of the majority of participants, who were male.

The simple main effect of sex of speaker on the acceptability ratings produced results very similar to the intelligibility ratings. Significant effects were found in cases with helicopter background noise and no background noise. Vocoder B exhibited a significant effect with jet background noise. As in the intelligibility ratings, the male speaker received higher ratings when background noise was present for both vocoders tested. In the absence of background noise, the female speaker received higher ratings for both vocoders as with the intelligibility analysis. This could indicate that there were some similarities between that which the participants scored intelligible and that which they scored acceptable.

4.1.4 Correlation Results

The results of the Pearson product moment coefficient of correlation between intelligibility and acceptability revealed a value of .37. This low value is indicative that these two subjective measures were relatively independent according to the aim of this study. This trend was evident within each level of the equipment independent variable, as well.

4.2 Analysis of Message Completion Test

Analysis of the objective responses of the message completion test indicated a lack of sensitivity to the intelligibility issue, although some significant results surfaced. Background noise influenced message completion results regardless of sex of speaker. However, the nature of that impact was somewhat different. For the male speaker, cases with helicopter background scored significantly higher than cases with jet background. For the female speaker, cases with propeller background noise scored significantly higher than cases with helicopter background. The relatively low scores for jet background is in direct concurrence with the subjective intelligibility scores presented previously.

The interaction of sex of speaker and equipment revealed significance only with the male voice condition. Under this condition, participants had higher message completion scores when clearances were presented through vocoder B. These results also concurred with those generally found in the subjective ratings portion of the test.

4.3 Analysis of Audio Preference Test

A look at the proportional responses for the Audio Preference Test revealed that, across all conditions, 86% of the responses favored vocoder B. Further, within each test condition, the lowest proportion for vocoder B selection was under conditions of the female speaker and helicopter background noise where 71% of the respondents selected vocoder B. The chi-squared analysis revealed that, in all cases studied, there was a clear preference for vocoder B. This fact, combined with the results of the previous subjective and objective evaluations, should leave no doubt concerning the issue of the superior performing vocoder.

Analysis of the preference rationale given by the participants revealed that the response reasons fell into 13 categories. For selections for the lesser favored vocoder A, the top three reasons were better intelligibility, less background noise, and a "guess." Selections for vocoder B revealed the top three reasons as better intelligibility, less background noise, and the voice quality being less garbled, the latter category being the fourth most prominent for vocoder A selection. The reasons for selection of each of the vocoders were quite similar and reflect those qualities most important to the air traffic controller participants.

5. Conclusions

This study has provided insight into the potential use of 4.8 kbps vocoders in the ATC system. The independent variables involved included sex of speaker, background noise, and equipment.

The dependent measures included intelligibility ratings, acceptability ratings, objective intelligibility scores, and a forced choice vocoder preference measure.

The results of the subjective analysis indicated interactions of the independent variables on ratings of intelligibility and acceptability. The intelligibility rating revealed that the analog radio control is more intelligible than vocoder B. However, vocoder B is more acceptable than analog radio. This may be a result of the static-free environment of the digital vocoders. Vocoder A rates lower in both areas. Jet background noise was the least favorable cockpit background noise for vocoder communications. This result is notable as jets comprise a large majority of the commercial aircraft fleet. Because no distinctive preference in sex of speaker arose, it is concluded that the sex of speaker has little effect on overall vocoder performance, although under specific conditions it produced significant effects.

The results of the Message Completion Test indicate that vocoder B is more intelligible than vocoder A for male speakers, still the majority in aviation today. This lent credibility to the results of the subjective ratings analysis. The lack of sensitivity of this test to the intelligibility issue prevented any comparison of analog radio to vocoders. Jet background noise affects the intelligibility of the vocoders in the most negative manner. Propeller background noise appears to pose the least threat to vocoder intelligibility. The objective results support the conclusion that sex of speaker has little effect on overall vocoder performance.

The results of the Audio Preference Test revealed a clear preference for vocoder B. The primary reasons for this selection were superior intelligibility and reduced background noise.

References

Child, J., Cleve, R., & Grable, M. (1989). *Evaluation of low data rate CODECS for air traffic control applications* (DOT/FAA/CT-TN89/13). Atlantic City, NJ: DOT/FAA Technical Center.

Crowe, D. P. (1988). *Selection of voice codec for the Aeronautical Satellite Service, AEEC subjective test report*. Ipswich, UK: British Telecom Research Labs.

Dehel, T., Grable, M., & Child, J. (1989). *Phase II testing and evaluation of low data rate voice CODEC equipment* (DOT/FAA/CT-TN89/49). Atlantic City, NJ: DOT/FAA Technical Center.

Dynastat. (1995). *Speech intelligibility evaluation at Dynastat*. Austin, TX: Dynastat, Inc.

Federal Aviation Administration. (1989). *Cockpit noise and speech interference between crewmembers* (DOT/FAA AC 20-133). Washington, D.C.

Fike, J. L., & Friend, J. E. (1983). *Understanding telephone electronics.* Ft. Worth: Texas Instruments Corporation.

Grable, M. (1990). *Phase III CODEC test plan* (DOT/FAA/CT-TN90/16). Atlantic City, NJ: DOT/FAA Technical Center.

Hart, S. D., (1988). Helicopter human factors. In Weiner, E. L., and Nagel, D. (Eds.). *Human factors in aviation*. San Diego: Academic Press, Inc.

Kemp, D. P., Sueda, R. A., & Tremain, T. E. (1989). *An evaluation of 4800 bps voice coders*. U.S. Department of Defense, International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

Miller, G., Heise, G., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology, 41*, 329-335.

Panzer, I. L., Sharpley, A. D., & Voiers, W. D. (1993). A comparison of subjective methods for evaluating speech quality, in *Speech and audio coding for wireless and network Applications*. Norwell, MA: Kluwer Academic Publishers.

Pickens, R. A. (1996). *Evaluation of vocoders in an aviation environment* (SC165/WG1-WP/193). Washington, DC: National Business Aircraft Association.

Rodgers, M. D. (1995). *A human factors evaluation of the operational demonstration flight inspection aircraft.* Oklahoma City, OK: Federal Aviation Administration Civil Aeromedical Institute.

Sanders, M. S., & McCormick, E. J. (1987). *Human factors in engineering and design.* New York: McGraw-Hill Book Company.

Tobias, J. V., (1968a). *Cockpit noise intensity: fifteen single-engine light aircraft.* Oklahoma City, OK: Federal Aviation Administration Office of Aviation Medicine.

Tobias, J. V., (1968b). *Cockpit noise intensity: eleven twin engine light aircraft.* Oklahoma City, OK: Federal Aviation Administration Office of Aviation Medicine.

Tremain, T. E., & Collura, J. S. (1988). *A comparison of five 16Kbps voice coding algorithms.* Washington, DC: U.S. Department of Defense, submitted for the International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

Tremain, T. E., Kemp, D. P., Collura, J. S., & Kohler, M. A. (1993). *Evaluation of low rate speech coders for HF.* Washington, DC: U.S. Department of Defense, submitted for the International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

Troll, N. L. (1989). *Assessment of voice coders for ATC/Pilot voice communications via satellite digital communication channels* (CAA89004). London: Civil Aviation Authority.

Voiers, W. D. (1983). Evaluating processed speech using the diagnostic rhyme test. *Speech Technology, 20*, pp. 30-39.

Warren, J. (1996). *Preliminary cockpit noise measurements* (Project 1705A). Atlantic City, NJ: DOT/FAA Technical Center.

Acronyms

ANOVA	Analysis of Variance
APT	Audio Preference Test
ARTCC	Air Route Traffic Control Center
ATC	Air Traffic Control
ATCS	Air Traffic Control Specialist
ATCT	Air Traffic Control Tower
BER	Bit Error Rate
DAM	Diagnostic Acceptability Measure
DF	Degrees of Freedom
DRT	Diagnostic Rhyme Test
FPL	Full Performance Level
HSD	Honestly Significant Difference
LPC	Linear Predictive Coding
OSHA	Occupational Safety and Health Administration
SATCS	Supervisory Air Traffic Control Specialist
SD	Standard Deviation
SME	Subject Matter Expert
SS	Sum of Squares
TRACON	Terminal Radar Approach Control
VHF	Very High Frequency

Appendix A
Test Samples

A.1. Subjective Ratings

INSTRUCTIONS: You will hear a short voice message either from a terminal or en route environment under a variety of background noise conditions. For advisories to airmen, you are asked to play the role of the pilot in command. For each of the phrases, you are asked to rate both the intelligibility and the acceptability on an 8-point scale according to the criteria defined below.

Intelligibility:

- *Ability to understand what was said in the message.*

Scale:

>-----*poor*----------*excellent*---->
1 2 3 4 5 6 7 8

Defined:

poor	Could not understand anything that was said during the transmission.
excellent	Understood everything that was relayed during the transmission precisely.

Acceptability:

- *Quality of the message: annoying, pleasant.*
- *Effort required to understand the message: easy, burdensome.*
- *Potential influence of the background noise: buzzing, hissing, etc.*

Scale:

>-----*poor*----------*excellent*---->
1 2 3 4 5 6 7 8

Defined:

poor	would be terribly annoying, frustrating, or unpleasant to hear.
excellent	excellent signal quality, a clear signal that would be pleasant to hear.

After hearing the message you are asked to select your intelligibility and acceptability ratings by circling the appropriate number. When you have finished your ratings, please look up at the test monitor so that he/she may proceed to the next audio phrase.

BEGIN

Phrase A1: ***"JETLINK TWENTY-ONE, LEFT HEADING OF TWO FIVE ZERO AND DESCENDING TO ONE THREE THOUSAND."***

Intelligibility

>-----*poor*----------*excellent*---->

1 2 3 4 5 6 7 8

Acceptability

>-----*poor*----------*excellent*---->

1 2 3 4 5 6 7 8

Phrase A2: ***"CLEVELAND, STING SIXTY-SEVEN WITH YOU AT FLIGHT LEVEL FOUR FIVE ZERO, REQUESTING DIRECT TO OFFUTT."***

Intelligibility

>-----*poor*----------*excellent*---->

1 2 3 4 5 6 7 8

Acceptability

>-----*poor*----------*excellent*---->

1 2 3 4 5 6 7 8

Phrase A3: ***"WE HAVE THE TRAFFIC IN SIGHT AND REQUESTING THE VISUAL APPROACH, COBRA SIX ONE FIVE."***

Intelligibility

>-----*poor*----------*excellent*---->

1 2 3 4 5 6 7 8

Acceptability

>-----*poor*----------*excellent*---->

1 2 3 4 5 6 7 8

A.2 Message Completion Test

INSTRUCTIONS: You will hear a phrase number followed by a short voice message either from a terminal or en route environment. For advisories to airmen, you are asked to play the role of the pilot in command. Each audio message you hear will correspond to a numbered phrase. The phrases will not be presented in the order in which they appear on the page. Using the spoken phrase number, you are asked to go to that phrase and print legibly in the space provided the missing parts which will make the statement complete. You may use any standard controller abbreviations. If you cannot recall a part of the message either guess or leave that space in the statement blank. When you have finished your response, please look up at the test monitor so that he may proceed to the next audio phrase. Please note that you may be asked to turn the page before all the phrases are completed.

BEGIN

Phrase A1: ***"TWA 127 DESCEND AND MAINTAIN FL190, TRAFFIC _____ O'CLOCK AND _____ MILES SOUTHWEST BOUND AT FLIGHT LEVEL _____."***

Phrase A2: ***"CENTER, USAIR _____, ANY REPORTS ON THE RIDE AT FLIGHT LEVEL _____, WE'RE PICKING UP _____."***

Phrase A3: ***"AMERICAN _____ TURN LEFT HEADING _____, CLIMB AND MAINTAIN FLIGHT LEVEL _____."***

Phrase A4: ***"_____ 2341, SQUAWK CODE _____ AND IDENT, EXPECT HIGHER ALTITUDE IN _____ MINUTES."***

Phrase A5: ***"WILMINGTON TOWER, MOONEY _____ APPROXIMATELY _____, MILES SOUTHWEST OF AIRPORT WITH _____."***

A.3 Audio Preference Test

INSTRUCTIONS: You will hear a voice message either from a terminal or en route environment. For advisories to airmen, you are asked to play the part of the pilot receiving the communication. Each spoken message corresponds to a written phrase below by number. You will hear each message twice but in different audio formats. After the second presentation, select the presentation format you would prefer in an ATC environment by circling either "FIRST" or "SECOND" following that phrase. Use the space provided underneath each clearance to tell, in your own words, what aspects of the communication led to your preferences.

BEGIN

Phrase	*Audio Preference*	
A1. "CLEARED THROUGH CHARLIE SURFACE AREA NORTHWEST OF CLEVELAND AIRPORT, WILL MAINTAIN SPECIAL VFR CONDITIONS AT OR BELOW FIVE THOUSAND, SIKORSKY THREE FOUR TANGO ROGER"	**FIRST**	**SECOND**
Reason(s):		

A2. "SIX MILES FROM FINAL APPROACH FIX, TURN RIGHT HEADING TWO THREE ZERO, MAINTAIN FOUR THOUSAND UNTIL ESTABLISHED ON THE LOCALIZER, CLEARED I-L-S RUNWAY TWO ZERO APPROACH, NOVEMBER THREE FOUR SEVEN KILO PAPA ROGER"	**FIRST**	**SECOND**
Reason(s):		

A3. "TURN RIGHT NEXT TAXIWAY, CROSS TAXIWAY BRAVO, HOLD SHORT TAXIWAY CHARLIE, WILL CONTACT GROUND ONE TWO ONE POINT SEVEN, CHEROKEE EIGHT ZERO TWO SEVEN LIMA ROGER"	**FIRST**	**SECOND**
Reason(s):		

Appendix B
Background Questionnaire

1) Please indicate your sex.

❑ Male ❑ Female

2) What is your age, in years?

_____ years

3) How many years have you actively controlled traffic?

_____ year(s)

4) How many months in the past year have you actively controlled traffic?

_____ month(s)

5) What is your current position as an air traffic controller?

❑ Developmental ❑ Full Performance Level ❑ Other

6) If you are a Full Performance Level controller, how long have you been at that level?

_____ year(s)

7) In which ATC environment do you currently work?

❑ En Route ❑ TRACON ❑ Tower Cab ❑ TRACON and Tower Cab